NATIONAL BUREAU OF ECONOMIC RESEARCH, INC.

SUMMER INSTITUTE 2012

Industrial Organization Program Meeting

Behavioral Economics: Glenn Ellison, Organizer

Digitization/IO: Susan Athey, Nick Bloom, Erik Brynjolfsson, Shane Greenstein, and Hal Varian, Organizers

IO General Session: Judy Chevalier and Ariel Pakes, Organizers

July 20-21, 2012

Royal Sonesta Hotel
Parkview Room
40 Edwin H. Land Boulevard
Cambridge, Massachusetts

PROGRAM

Friday, July 20:

8:30 am Coffee and Pastries

9:00 am Mark Armstrong, Oxford University
Yongmin Chen, University of Colorado
Discount Pricing

Discussant: Glenn Ellison, Massachusetts Institute of Technology and NBER

9:45 am Michael Grubb, Massachusetts Institute of Technology and NBER
Matthew Osborne, Bureau of Economic Analysis
Cellular Service Demand: Biased Beliefs, Learning, and Bill Shock

Discussant: Andrew Sweeting, Duke University and NBER

10:50 am Break

11:05 am Fabian Duarte, RAND Corporation
Justine Hastings, Brown University and NBER
Fettered Consumers and Sophisticated Firms: Evidence from Mexico's Privatized Social Security Market

Discussant: Phillip Leslie, University of California at Los Angeles and NBER

11:50 pm Meghan Busse, Northwestern University and NBER
Devin Pope, University of Chicago and NBER
Jaren Pope, Brigham Young University
Jorge Silva-Risso, University of California at Riverside
Projection Bias in the Car and Housing Markets

Discussant: Chad Syverson, University of Chicago and NBER

12:30 pm Lunch

Joint session with the Economics of IT and Digitization Workshop

11:45 am Alternate Lunch and Panel:
(Ballroom, West Tower)

1:30 pm Elisa Celis, University of Washington
Gregory Lewis, Harvard University and NBER
Markus Mobius, Iowa State University and NBER
Hamid Nazerzadeh, University of Southern California
Buy-it-Now or Take-a-Chance: A Mechanism for Real-Time Price Discrimination

2:15 pm Joerg Claussen, Ifo Institute, University of Munich
Tobias Kretschmer, LMU Munich
Philip Mayrhofer, CDTM, LMU & TU Munich
Incentives for Quality over Time – The Case of Facebook Applications

3:00 pm Break

3:30 pm Liran Einav, Stanford University and NBER
Chiara Farronato, Stanford University
Jonathan D. Levin, Stanford University and NBER
Neel Sundaresan, eBay Research Labs
Sales Mechanisms in Online Markets: What Happened to Online Auctions?

Paper under pre-release review; to be added later

4:14 pm Dina Mayzlin, Yale University
Yaniv Dover, Yale University
Judith A. Chevalier, Yale University and NBER
Promotional Reviews: An Empirical Investigation of Online Review Manipulation

5:00 pm Adjourn

6:00 pm IO Program Dinner, Legal Sea Foods at Kendall Square

Saturday, July 21

IO General Program Meeting

8:30 am Coffee and pastries

9:00 am Alon Eizenberg, Hebrew University of Jerusalem
Alberto Salvo, Northwestern University
Grab them Before they Go Generic: Habit Formation and the Emerging Middle Class

Discussant: Matthew Gentzkow, University of Chicago and NBER

10:00 am Break

10:15 am Gautam Gowrisankaran, University of Arizona and NBER
Robert Town, University of Pennsylvania and NBER
Aviv Nevo, Northwestern University and NBER
Keith Brand, US Federal Trade Commission
Christopher J. Garmon, Federal Trade Commission
Mergers When Prices Are Negotiated: Evidence from the Hospital Industry

Paper under pre-release review, to be added later

Discussant: Robin Lee, New York University

11:15 am Break

11:30 am Kerem A. Cosar, University of Chicago
Paul Grieco, Pennsylvania State University
Felix Tintelnot, Pennsylvania State University
Borders, Geography, and Oligopoly: Evidence from the Wind Turbine Industry

Discussant: Jan De Loecker, Princeton University and NBER

12:30 pm Lunch

1:30 pm Frank A. Wolak, Stanford University and NBER
Measuring the Competitiveness Benefits of a Transmission Investment Policy: The Case of the Alberta Electricity Market

Discussant: Steven Puller, Texas A&M University and NBER

2:30 pm Break

2:45 pm Sang Won Kim, Columbia University
Marcelo Olivares, Columbia University
Gabriel Weintraub, Columbia University
Measuring the Performance of Large-Scale Combinatorial Auctions: A Structural Estimation Approach

Discussant: Gregory Lewis, Harvard University and NBER

3:45 pm Break

4:00 pm Keith Ericson, Boston University
Amanda Starc, University of Pennsylvania
Age-Based Heterogeneity and Pricing Regulation on the Massachusetts Health Insurance Exchange

Discussant: Benjamin Handel, University of California at Berkeley and NBER

5:00 pm Adjourn

Discount Pricing*

Mark Armstrong
Department of Economics
University of Oxford

Yongmin Chen
Department of Economics
University of Colorado at Boulder

May 2012

Abstract

This paper investigates *discount pricing*, the common marketing practice whereby a price is listed as a discount from an earlier, or regular, price. We discuss two reasons why a discounted price—as opposed to a merely low price—can make a rational consumer more willing to purchase the item. First, the information that the product was initially sold at a high price can indicate the product is high quality. Second, a discounted price can signal that the product is an unusual bargain, and there is little point searching for lower prices. We also discuss a behavioral model in which consumers have an intrinsic preference for paying a below-average price. Here, a seller has an incentive to offer different prices to identical consumers, so that a proportion of its consumers enjoy a bargain. We discuss in each framework when a seller has an incentive to offer false discounts, in which the reference price is exaggerated.

Keywords: Reference dependence, price discounts, sales tactics, false advertising.

1 Introduction

In his account of sales practices, Cialdini (2001, page 12) writes about

> the Drubeck brothers, Sid and Harry, who owned a men's tailor shop [...] in the 1930s. Whenever Sid had a new customer trying on suits in front of the shop's three-sided mirror, he would admit to a hearing problem and repeatedly request that the man speak more loudly to him. Once the customer had found a suit he liked and asked for the price, Sid would call to his brother, the head tailor, at the back of the room, 'Harry, how much for this suit?' Looking up from his work—and greatly exaggerating the suit's true price—Harry would call back, 'For that beautiful, all wool suit, forty-two dollars.' Pretending not

*We are grateful to David Gill, Salar Jahedi, Andrew Rhodes, Mike Riordan, Rani Spiegler, John Vickers and Jidong Zhou for helpful discussions.

> to have heard and cupping his hand to his ear, Sid would ask again. Once more Harry would reply, 'Forty-two dollars.' At this point, Sid would turn to the customer and report, 'He says twenty-two dollars.' Many a man would hurry to buy the suit and scramble out of the shop with his [...] bargain before poor Sid discovered the 'mistake'.

As this anecdote suggests, consumers are more likely to buy an item if they perceive it to be a bargain. This is easily understood when the consumer is given an *accidental* discount, as occurs for instance if she sees that the product she wants has been given the wrong price tag. If the product's genuine price—which reflects its cost, quality and/or competitive environment—is \$42, but by chance the consumer can get the product for \$22, this represents genuine value-for-money and will make the consumer more inclined to purchase. This rational response to an accidental discount is exploited by the Drubecks' fraudulent sales tactics.

What is more of a challenge is to explain why consumers might care about receiving a *deliberate* discount from a seller, as opposed simply to obtaining a low price. For instance, a consumer may be more likely to buy a jacket priced at \$100 accompanied by a sign which reads "50% of its previous price" than he/she would be if the price were merely stated as \$100. Alternatively, a retailer might claim its price was \$100 even though the "manufacturer's recommended price" was \$200. Despite its prevalence, this pricing practice—which we term *discount pricing*—has apparently received little economic analysis. In the literature on sales (for instance, Lazear 1986), consumers care only about the price level, and whether a low price is framed as a discount off a higher price plays no role. In this paper, we explore the economics of discount pricing, focussing on the potential information content of a discount and its strategic implications. Our analysis is developed in two models that suggest different reasons why rational consumers care about discounts, as well as in a third model with behavioural consumers.

First, in section 2, uninformed consumers rationally take a monopoly seller's initial price as a signal of its choice of quality, and so are willing to pay more for the product when they observe the initial price was high. The firm sells its product to two groups of consumers, one of which can accurately determine the product's quality while the other group, the casual buyers, cannot. The monopolist can price discriminate between the two consumer groups using inter-temporal pricing, and the second group can use the price offered to the first group, when they observe it, as an indicator of quality. In this framework, it is more likely that the firm has an incentive to supply a high-quality product when casual buyers

can observe its initial price. Thus, the firm's ability to write "was $200, now $100", if credible, may induce it to provide a high-quality item.

In the second model, presented in section 3, the knowledge that a product is offered at a discounted price induces consumers to buy immediately rather than investigate a rival's price. Two firms compete to sell to consumers, and either firm offers one of two prices: a full price or a sale price. (Price variation is generated by exogenous demand variation.) When a product is offered on sale, a consumer buys immediately even if that price is relatively high, and so a consumer cares about whether a discount is offered rather than the level of the actual price. If a consumer is given no credible information about whether the current price is discounted or not, she must judge how likely it is that the next price will be higher, given the current price, and buy accordingly. This inability to fine-tune her search process can cause welfare losses.

In our third model, in section 4, consumers intrinsically care about getting a bargain. Studies in behavioral economics (discussed shortly) have focused on how "reference prices", which can sometimes be manipulated by a seller's marketing activities, affect purchase decisions. In our model, a bargain is a price below the firm's average offered price. If consumers observe the firm's prices to all consumers, the firm responds to the "demand for bargains" by offering distinct prices to otherwise identical consumers. If the demand curve is concave, the firm follows a simple "high-low" pricing strategy with just two prices, a full price and a sale price. If instead consumers see only their own price, but hold equilibrium beliefs about the average price, the firm again has an incentive to pursue a high-low policy, but one with lower prices relative to when consumers see the prices offered to all consumers. When its prices are secret, the firm has a greater incentive to undercut its anticipated average price to some consumers, since others do not see this price cut and cannot react to it.

If, for whatever reason, consumers care about getting a discount, a seller may have an incentive to exploit this by making false claims about its previous or regular price. The outcome when these deceptive marketing tactics are used depends on the "savviness" of consumers. If consumers are aware that sellers are able to misrepresent their reference price without penalty, they will simply regard such sale signs as puffery and pay them no attention. The result is that a potentially useful channel of information is absent. However, if instead consumers are more gullible and believe a firm's false claims (when such claims are plausible), the outcome is worse, as these consumers may be induced to pay more for the product than they would otherwise.

The media regularly features stories in which a seller's claimed discounts are alleged to be fictitious. For instance, a supermarket's heavily advertised 15% average price reduction may have been preceded by an unadvertised gradual price rise cancelling out the reduction. In Britain, a legal case involved the "Officers Club" chain of clothing stores, where it was alleged that only a tiny share of sales were made at the regular price and the great majority of items were sold at "70% off" this supposed regular price.[1] Several jurisdictions have rules in place to combat false discounting.[2] In the United States, the Federal Trade Commission's *Guides Against Deceptive Pricing* (para. 233.1) distinguishes between genuine and fictitious discounts. For instance, "where an artificial, inflated price was established for the purpose of enabling the subsequent offer of a large reduction - the 'bargain' being advertised is a false one; the purchaser is not receiving the unusual value he expects. In such a case, the 'reduced' price is, in reality, probably just the seller's regular price."[3]

There are a number of earlier contributions which discuss issues related to our models. Our first model, where an initial price of a product signals its quality, builds on a large literature which studies how (current) price can signal quality. For instance, Bagwell and Riordan (1991) present a model where a firm has private information about the exogenous quality of its product. They find that high and declining prices signal high product quality: the firm distorts its price above the full-information level in order to signal high quality, and, as more consumers become informed, there is less price distortion in later periods. While their motivation is different from ours and their insights are derived mainly in a setting where the firm's current price signals quality, they also consider an extension where consumers can observe the firm's past price. In this case, the firm's prices may be more distorted in period 1 but less distorted in period 2, compared to when past price is not observed, and they find that the high-quality firm has an incentive to reveal past price information to uninformed consumers. Thus, when a firm makes sequential sales of a product, the exogenous quality of which is the firm's private information, a policy that

[1] The England and Wales High Court (Chancery Division) found that the seller engaged in "misleading advertising". See details in the judgement of 26 May 2005 of Justice Etherton of the case between the Office of Fair Trading and The Officers Club Ltd., www.bailii.org/ew/cases/EWHC/Ch/2005/1080.html. For instance, in paragraph 16 of this judgement, it states that between 1 September 2002 and 28 June 2003 only 0.15% of the total number of items sold in the chain of stores were at the "full price". The judgement also discusses similar cases in other countries, such as *Colorado vs. May Department Stores* in the United States (para. 59), and *Commissioner of Competition vs. Sears Canada Inc.* in Canada (para. 63).

[2] Some jurisdictions also have policies to prevent permanent sales by requiring all sales to occur on stipulated dates. Thus the winter sales in Paris in 2012 had to take place between 11 January and 14 February.

[3] This document can be downloaded from www.ftc.gov/bcp/guides/decptprc.htm.

bans false discounts would boost profit.

Muris (1991, section IIIC) and Rubin (2008, section III) discuss how the FTC has ceased fighting fictitious pricing cases in recent years, in part because it was often rival sellers—not consumers—who used the FTC's *Guides* to prevent a firm's heavy discounting, and in part because of a perception that any focus on price was potentially pro-competitive. However, our second model in section 3 suggests that complaints by rivals about a firm's false sales can have a procompetitive motive: false discounts discourage consumers from investigating rival offers and deprive the rivals of opportunity to compete effectively. In these settings, preventing false discounting can lead to more effective competition.

Models and experiments from psychology and behavioral economics offer a number of insights on the use of discount pricing.[4] Thaler (1985) proposes a model of consumer behaviour in which the context of a transaction matters to a consumer as well as the transaction itself. One implication of this theory is that firms can profit from a high "suggested retail price", which serves as a reference price, and a lower selling price may then provide consumers with a "transaction utility". Bordalo, Gennaioli and Shleifer (2012) develop a model of salience in consumer decision making, which they use to explain a number of perplexing phenomena. Their analysis suggests that, by raising consumers' valuation of quality through salience, firms can benefit from "misleading sales"—artificially inflating the regular price and simultaneously offering a generous discount. Jahedi (2011) experimentally investigates a kind of "bargain" which we do not study in this paper, where a seller offers two units of its product for little more than the price of one unit. He shows how consumers are less likely to buy two units when faced with the choice from {buy nothing, buy two units for \$1} than they are when faced with the larger choice set {buy nothing, buy one unit for \$0.97, buy two units for \$1}. Jahedi designs the experiments so that subjects know that prices have no signaling role (such as the signaling roles we analyze in our first two models), and deduces that his subjects have an intrinsic "taste for bargains".

Our third model is a model with consumer reference dependence, where consumers also have a taste for bargains. Spiegler (2011a, section 9.4.2) briefly outlines a related model, although his construction perhaps uses implausibly high prices (higher than any consumer's raw valuation for the product). Most existing models of consumer reference dependence

[4]Experimental evidence that consumers are influenced by false sales is discussed by Urbany, Bearden and Weilbaker (1988). They also found more generally that an advertised reference price—plausible or exaggerated—raised consumers' estimates of the firm's regular price and the perceived offer value, and reduced consumer search for other sellers.

focus instead on loss-aversion, where a consumer's propensity to buy falls when offered a price above her reference point. See Heidhues and Kőszegi (2005), Spiegler (2011b), Puppe and Rosenkranz (2011) and Zhou (2011) for models involving consumer loss aversion. Much of this literature finds that loss aversion makes a firm's prices more rigid, for instance in response to cost variation, than would be the case in a "standard" model. By contrast, when consumers are bargain-loving, we show that a firm is *more* inclined to vary its prices than otherwise.

2 Initial Price as Signal of Product Quality

In this section we modify a standard static model of quality choice so that the firm sells over time.[5] Specifically, a monopolist supplies a product over two periods, with its price in period $t = 1, 2$ denoted p_t, and chooses its quality *ex ante* which is then fixed for the two periods. The firm can choose one of two quality levels, L and H, and it has constant unit cost c_i if it chooses quality $i = L, H$. All consumers have unit demand. For simplicity, suppose the firm aims to maximize the sum of profits in the two periods.

A fraction σ of consumers are *keen* and particularly interested in the product: they can discern the product's quality, and they are impatient and wish to buy only in period 1. Their valuation is v_i for the product when its quality is $i = L, H$. The remaining $1 - \sigma$ consumers are *casual* buyers: they cannot directly observe quality and buy for simplicity only in period 2. (Little of substance in the analysis would be affected if some casual buyers also purchased in the first period.) Their valuation for the product is θv_i when quality is $i = L, H$, where the parameter $0 < \theta \leq 1$ reflects the plausible situation where casual buyers have a lower willingness-to-pay for the item. To avoid discussing sub-cases involving non-supply, we assume that

$$\theta v_L > c_H \tag{1}$$

so that the high-quality product can profitably be sold even to casual buyers who think quality is low. We also assume that providing the high-quality product is socially efficient, so that

$$[\sigma + (1 - \sigma)\theta] \Delta_v > \Delta_c \ , \tag{2}$$

where $\Delta_v \equiv v_H - v_L$ and $\Delta_c \equiv c_H - c_L$.

[5] This static model is taken from Tirole (1988, section 2.3.1.1), which itself incorporates elements from a number of earlier contributions.

We study market equilibrium under alternative information assumptions. A consumer buys the item if the price is no higher than her willingness-to-pay, which depends on observed (if the consumer is keen) or anticipated (if casual) product quality. The firm's strategy consists of its choice of quality and its two prices. In equilibrium the firm's strategy is optimal given consumer buying behaviour, while the expectations of product quality by casual buyers, which may depend on observed prices, are consistent with the firm's strategy.

Consider first the case where the casual buyers do not observe the firm's initial price. A casual buyer's anticipated quality might depend on the period-2 price. However, all that matters for the firm is the *maximum* price, say P, which induces a casual buyer to buy the product. (If the firm is going to sell to casual buyers it should set the highest possible price, regardless of its chosen quality.) Clearly, we have $\theta v_L \leq P \leq \theta v_H$, since the value of the item to the casual buyers is known to lie between these extremes. From (1), it is profitable to sell to these casual buyers, regardless of their beliefs about quality. Thus, given P, the firm's profit if it chooses to supply the high-quality product is

$$\sigma(v_H - c_H) + (1 - \sigma)(P - c_H) \ ,$$

while its profit if it supplies the low-quality product is

$$\sigma(v_L - c_L) + (1 - \sigma)(P - c_L) \ .$$

Comparing these two profits, we see that if

$$\sigma > \frac{\Delta_c}{\Delta_v} \ , \tag{3}$$

the unique equilibrium is for the firm to provide a high-quality product, and the firm's prices fully extract consumer surplus so that $p_1 = v_H$ and $p_2 = \theta v_H$. Thus, if the fraction of informed buyers is large enough, the firm makes more profit by serving these buyers with their preferred product than by supplying a low-cost product to all consumers. By contrast, if $\sigma < \Delta_c/\Delta_v$ the unique equilibrium is to provide a low-quality product, and prices are $p_1 = v_L$ and $p_2 = \theta v_L$. We summarize this discussion as:

Lemma 1 *Suppose that casual buyers cannot observe the firm's initial price. If the fraction of keen buyers is large enough that (3) is satisfied, the unique rational expectations equilibrium is for the firm to supply a high-quality product, and to choose prices which fully extract consumer surplus (i.e., $p_1 = v_H, p_2 = \theta v_H$). If the fraction of keen buyers is small enough that (3) is strictly violated, the unique rational expectations equilibrium is for the*

firm to supply a low-quality product, and to choose prices which fully extract consumer surplus (i.e., $p_1 = v_L, p_2 = \theta v_L$).

Consider next the case where casual buyers do observe the initial price. For instance, they see a price label which truthfully states "was \$200, now \$100". A similar argument to that used for Lemma 1 establishes that when (3) holds, providing high quality is the unique equilibrium. But now, even if (3) fails, high quality can be supported in equilibrium. Specifically, suppose the firm chooses a particular initial price p_1 such that $v_L < p_1 \leq v_H$. Suppose given p_1 that the maximum price which induces the casual buyers to buy is P, where as before P lies in the range $\theta v_L \leq P \leq \theta v_H$. Then the firm's profit if it supplies a high-quality product is

$$\sigma(p_1 - c_H) + (1 - \sigma)(P - c_H) \ ,$$

while its profit if it provides a low-quality product is

$$(1 - \sigma)(P - c_L) \ .$$

(For this last expression, note that the firm does not sell to the informed buyers since $v_L < p_1$.) Thus, supplying a high-quality product is more profitable if

$$\sigma(p_1 - c_L) > \Delta_c \ . \tag{4}$$

In particular, we see that a higher initial price makes it more likely that offering a high-quality product is profitable, and in this sense a high initial price acts as a signal to casual buyers that quality is high. The reason is that a high initial price makes deviating to low quality more costly for the firm: if it deviates to low quality, it must forego serving the keen (informed) buyers and serving these buyers is more profitable with a higher initial price. Setting $p_1 = v_H$ in (4) implies that first-best profit—where the firm supplies a high-quality product and chooses prices $p_1 = v_H$ and $p_2 = \theta v_H$— is feasible if $\sigma(v_H - c_L) > \Delta_c$, i.e., if

$$\sigma > \frac{\Delta_c}{\Delta_v + [v_L - c_L]} \ . \tag{5}$$

If this condition does not hold, there is *no* initial price which could convince casual buyers that quality is high. In this case the firm supplies a low-quality item and fully extracts the resulting consumer surplus.

Since condition (5) is less stringent than (3), we deduce that efficient quality provision is easier to achieve when the initial price is observed by casual buyers. When (5) holds but (3) does not, there is another equilibrium with low quality. As is usual in signaling games,

this multiplicity of equilibrium is due to the arbitrariness of beliefs off the equilibrium path. For clear-cut statements in the rest of this section, we assume that beliefs off the equilibrium path satisfy the "forward induction" refinement: when seeing a price off the equilibrium path, casual buyers reason what quality the firm could have rationally chosen given this price; if it is always optimal for the firm to choose q, then their belief is that quality is q. Then, when (5) holds, high quality is the unique equilibrium.[6]

We summarize this discussion as:

Lemma 2 *Suppose that casual buyers can observe the firm's initial price. If the fraction of keen buyers is large enough that (5) is satisfied, the unique rational expectations equilibrium is for the firm to supply a high-quality product, and to choose prices which fully extract consumer surplus (i.e., $p_1 = v_H, p_2 = \theta v_H$). If the fraction of keen buyers is small enough that (5) is strictly violated, the unique rational expectations equilibrium is for the firm to supply a low-quality product, and to choose prices which fully extract consumer surplus (i.e., $p_1 = v_L, p_2 = \theta v_L$).*

If the firm can credibly reveal its initial price to casual buyers, then when the fraction of keen buyers lies in the range

$$\frac{\Delta_c}{\Delta_v + [v_L - c_L]} < \sigma < \frac{\Delta_c}{\Delta_v} \tag{6}$$

the firm will wish to do so. (When the fraction lies outside this range, communicating its initial price to casual buyers has no impact, as anticipated quality cannot be affected by the firm's initial choice of price.) Welfare—which equals profit in this setting with full extraction of consumer surplus—also rises in this case.

We summarize the discussion as:

Proposition 1 *Relative to a setting where casual buyers cannot observe the initial price, if the firm can credibly communicate the initial price to casual buyers, this weakly (strictly if condition (6) holds) increases product quality, profit and welfare.*

Now consider the scenario in which the firm is able to make any claim—true or false—about its initial price. If casual buyers are aware that the firm can make false claims about its discount without penalty, they will "discount the discount" and behave just as if they

[6] At the potential low-quality equilibrium (with $p_1 = v_L$ and $p_2 = \theta v_L$), consider a deviation to high quality with $p_1 = v_H$ and $p_2 = \theta v_H$. Since with $p_1 = v_H$ it is always optimal for the firm to choose quality H, regardless of what P is, the forward induction refinement implies that the casual buyers must believe that the firm has chosen H upon seeing $p_1 = v_H$. This eliminates the low-quality equilibrium.

do not observe the initial price. When the fraction of keen buyers lies in the range (6), a policy which prevents firms making false claims about discounts will induce the firm to switch from offering a low-quality to a high-quality product, which will boost profit and welfare. The policy opens up a useful channel of information to otherwise uninformed buyers. In particular, if the casual buyers are savvy in this manner, the firm will *welcome* a policy which forbids it from making fictitious discount claims.

However, casual buyers might instead be "gullible" and believe the firm's claims about its initial price when such claims are plausible. For instance, they might mistakenly think that effective consumer policy is already in place to prevent misleading price claims.[7] If the fraction of keen consumers lies in the range (6), then faced with these more gullible casual buyers the firm would not switch to offering a high-quality product. Instead, the firm would produce a low-quality product, actually offer the initial price $p_1 = v_L$ to the keen buyers, but claim to casual buyers that its initial price was $p_1 = v_H$, who can therefore be charged price $p_2 = \theta v_H$. The outcome is poor for casual buyers, who suffer negative consumer surplus. Thus, in the case with gullible consumers a policy which prevents misleading claims about initial prices not only ensures efficient quality choice (as was the case with savvy consumers), but now improves consumer welfare and reduces profit.

The idea that consumers care about a seller's initial price because it signals product quality can be applied to other settings. Consider, for instance, the following variant of Lazear's (1986) model of clearance sales. Suppose that the firm has only one unit of a product to sell, and that the quality of its product, denoted v, is exogenous, uncertain, and initially unobserved even by the firm itself.[8] In the first period, a keen consumer who observes v considers buying the product, and will buy if the initial price p_1 is below v. If he chooses not to buy the product, a casual consumer in the second period considers whether to buy. The casual buyer does not directly observe v, and bases her purchase decision on the expected value of v, conditional on the item not having sold in the first period. In this setting, total supply is limited, and when the casual buyer sees the item on sale in period 2, she knows that demand from the keen buyer was low. This causes her to lower her estimate of quality. But the information content of the event that the item ends up on sale is less when the initial price was high, as fewer informed consumers would have been willing to buy at a higher price. That is, expected quality, conditional on the item remaining unsold,

[7] In an environment where casual buyers do not observe initial prices, there is no difference between "savvy" and "gullible" consumers, and both make rational inferences about a firm's choice of quality.

[8] In the fashion context, for example, v might represent whether or not the product's colour or cut is fashionable that season, which is not something the firm knows in advance.

is higher when the initial price was higher.[9] Hence, initial price again acts as a signal of quality, albeit for a reason very different to that in our endogenous quality model. It can be shown, however, that in this setting firm profit is lower, and consumer surplus is higher, when the initial price is *not* observed. (We will obtain a similar result in our model presented in section 4.)

3 Discounts as a Signal to Buy Immediately

A second reason why consumers like a discounted price is because this may signal the price is unusually low, and they would do well to take advantage of it. This signal could potentially operate in two dimensions. In a monopoly context where the firm sets different prices over time, a discounted rather than full price might indicate the price is likely to go up, and the consumer should buy immediately rather than wait for a lower price. Alternatively, in a static oligopoly search context, a discounted price from one seller could indicate that rival prices are likely to be no lower, and there is little reason to investigate other sellers when search is costly. In this section we explore the latter possibility. (The dynamic monopoly model can be analyzed in a very similar manner.)

Before describing the analysis in detail, we point out that to investigate the question at hand we need a framework which is more complicated than standard models of search. As usual, we require a framework with price dispersion so that consumers sometimes have an incentive to search for a lower price. However, in order to discuss the impact of discounted prices, as opposed to merely low prices, we need the pattern of price dispersion itself to be uncertain from the consumer's point of view. For instance, if the consumer knew the potential prices were p_L and p_H, then if she first encounters p_L she knows the other price is either p_L or p_H and so does not benefit from additional information about whether the price is discounted.

In more detail, suppose two firms compete to sell a homogeneous product to consumers. The two firms sell repeatedly over time, although all consumers are short-lived and can buy only in their own period. A firm's price is either p_L or $p_H > p_L$ in each period with the probability of the latter being α, and price is independently realized in each period and across firms. We refer to p_L as the "sale" (or discounted) price and p_H as the "regular" (or full) price. The market parameters (p_L, p_H, α) are unchanging over time. Thus the regular and the sale prices are the same for both firms, although with probability $2\alpha(1-\alpha)$ one

[9] In Lazear's model, the second consumer is also well informed about v, and so does not care about the initial price.

firm runs a sale while its rival does not. For now we take the process of price determination to be exogenous. (The model will be "closed" in a particular way shortly.)

Suppose there are a number of "searchers" who are imperfectly informed about market prices. Specifically, they can travel to their local firm for free and see its price and, if desired, buy immediately from that firm (with an equal proportion of consumers local to each firm), but they need to incur a cost s_1 to travel to the second, to them more remote, firm and discover its price. Suppose a consumer can return to buy from her local firm after investigating the remote firm by incurring the further search cost s_2. Suppose prices are such that these searchers will always wish to buy the product from one firm or the other.

The ideal search rule for such a consumer, given known tariff parameters (p_L, p_H, α), is simple. If the consumer knows the local price is the sale price she will buy immediately, as the rival's price cannot be lower. If the consumer knows the local price is the full price, she may decide to investigate the rival's price in case it turns out to be discounted. If her local firm offers p_H, the risk-neutral consumer has an incentive to investigate the remote firm whenever $p_H \geq s_1 + \alpha p_H + (1-\alpha)p_L$, i.e., when the expected sale discount $(1-\alpha)(p_H - p_L)$ satisfies

$$(1-\alpha)(p_H - p_L) \geq s_1 \ . \tag{7}$$

If the local price is the full price, the consumer will nevertheless buy locally if (7) does not hold, as it is not worth incurring the search cost to obtain the small expected discount at the rival. A consumer will never return to buy from her local firm after travelling to the remote firm. This ideal stopping rule depends on whether the local product is offered on sale, and on the size and frequency of the sale discount, but not on price *levels*.

Now suppose a consumer is initially offered price p from her local firm, without any credible information about whether this price is discounted. She must then decide whether to buy immediately purely on the basis of the price level. Moreover, a consumer might sometimes return to buy locally after travelling to the remote firm, thus incurring a double search cost $s_1 + s_2$. Suppose that tariff parameters (p_L, p_H, α) are uncertain from the viewpoint of the consumer. A consumer conditions the distribution of rival's price $\tilde{p}$ on the local firm's price p, and a consumer who sees local price p will buy immediately if and only if

$$p \leq s_1 + \mathbb{E}[\min\{\tilde{p}, p + s_2\} \mid p] \ . \tag{8}$$

Here, the right-hand side is the expected expense involved if the consumer travels to the remote firm: the search cost s_1 is sunk, but then the consumer has the ability to buy from whichever supplier is cheaper (after taking the cost of returning to the local seller into

account). The search rule in (8) will in general be inefficient compared to the search rule when the consumer knows when the local price is the discounted price, and so we expect that credible information about discounts will benefit consumers.

To investigate in more detail, we specialize and close the model in the following manner. Here, the firms' price variation is generated by local demand shifts.[10] Specifically, suppose in each period there are also a number of "inert" consumers can buy only from their local firm (to which they can travel costlessly). These consumers have unit demand, and their valuation for the unit can take one of three values: V_L, V_M or V_H, where $0 < V_L < V_M < V_H$. The market operates in one of two states. In the first state, the possible valuations are $\{V_L, V_M\}$, and in any period and for either firm these two demand realizations are equally likely. In the other state, the possible valuations are $\{V_M, V_H\}$ and again these two demand realizations are equally likely. Each market state $\{V_L, V_M\}$ or $\{V_M, V_H\}$ is realized *ex ante* with equal probability. A firm knows which market state is realized, but in a given period does not observe its rival's local demand realization. Suppose each firm's production is costless.

The searchers are willing to pay up to V_H for a single unit and their search costs are $s_1 = s_2 = s$, where $0 < s < V_H - V_M$. (The condition $s < V_H - V_M$ will ensure that the consumer will return to buy locally if she discovers the remote price is higher.) The key feature of this set-up is that when a searcher knows that local demand is V_M, she does not know if the market state is $\{V_L, V_M\}$ or $\{V_M, V_H\}$. We will derive an equilibrium in which each firm sets its price to fully extract surplus from their inert consumers, i.e., a firm chooses $p = V_i$ when its realized local demand is V_i. Intuitively, this pricing behaviour is an equilibrium whenever the proportion of searchers is small enough, as then a firm's incentive to extract surplus from the inert consumers dominates the incentive to keep the searchers from investigating the rival firm.

For now, take as given this pricing rule by firms. What is the optimal search rule for the searchers? From (8), and given $s < V_H - V_M$, a consumer has an incentive to travel to the remote firm when the local price is $p = V_M$ (and then to travel back to the local firm if the remote price turns out to be $p = V_H$) if and only if

$$s \leq \tfrac{1}{5}(V_M - V_L) \ . \tag{9}$$

A consumer has an incentive to travel to the remote firm when the local price is $p = V_H$ if

[10] There are other ways to close the model. For instance, we might have inert consumers with a constant downward-sloping demand curve, and each firm has idiosyncratic shocks to its unit cost.

and only if

$$s \leq \tfrac{1}{2}(V_H - V_M) \ . \tag{10}$$

(Of course, a consumer will buy immediately if she is offered the lowest price V_L.) A complicating factor is that this search rule may not be monotonic; that is, a consumer might search on when she sees the intermediate price $p = V_M$ but not if she sees the highest price $p = V_H$. The reason is that in the latter case, the consumer knows that the low price $p = V_L$ is not a possibility, and it might be that this chance of the low price is what drives search incentives when $p = V_M$. This possibility is ruled out if condition (9) implies condition (10), i.e., if $\frac{1}{2}(V_H - V_M) \geq \frac{1}{5}(V_M - V_L)$. In particular, if the search cost is small enough that

$$s < \tfrac{1}{5}(V_M - V_L) \leq \tfrac{1}{2}(V_H - V_M) \ , \tag{11}$$

the optimal search rule is to buy immediately if the local price is $p = V_L$ and otherwise to travel to the remote firm. (If the local price is $p = V_M$ and the remote price is $p = V_H$, the consumer will then return to buy locally.) Of course, this search rule is inefficient, as when the market state is $\{V_M, V_H\}$ and the consumer is first offered price $p = V_M$, she travels to the remote firm even though the price cannot be lower there. Nevertheless, the consumer always buys the product at the cheapest price available.

The following result describes market equilibrium when consumers do not know whether their local price is discounted or not:

Lemma 3 *Suppose parameters satisfy (11). Provided the proportion of searchers in the consumer population is sufficiently small, the following strategies make up an equilibrium when searchers have no credible information about whether the local price is discounted: (i) each firm sets its price to extract surplus fully from their inert consumers, i.e., a firm chooses $p = V_i$ when its realized local demand is V_i, and (ii) searchers buy immediately if the local price satisfies $p \leq V_L$ and otherwise they travel to the remote firm.*

Proof. We have already shown that this search rule is optimal given the claimed price choice by firms. To see that firms optimally price in the stated way given this consumer search rule whenever the proportion of searchers is sufficiently small, argue as follows. Suppose the number of inert consumers is N and the number of searchers is n. Suppose for instance that the market state is (V_L, V_M) and a firm's demand realization is V_M. If the firm follows the stated strategy and sets price $p = V_M$, its expected profit is $V_M(\frac{1}{2}N + \frac{1}{4}n)$ since its $\frac{1}{2}N$ inert consumers will buy and the $\frac{1}{2}n$ searchers local to the rival firm will buy from it if the rival price is also V_M, which occurs with probability $\frac{1}{2}$. (The firm's local searchers

will never buy from it.) If the firm deviates to price $p = V_L$, its profit is $V_L(\frac{1}{2}N + \frac{3}{4}n)$, since now the firm's local searchers will buy from it as well. The latter profit is below the former when $\frac{n}{n+N}$ is small. Another potentially profitable deviation is to set price $p = V_M - s$, which will induce all searchers to buy from it in the event the rival price is $p = V_M$, and so generates profit $(V_M - s)(\frac{1}{2}N + \frac{1}{2}n)$. This is below $V_M(\frac{1}{2}N + \frac{1}{4}n)$ whenever the proportion of searchers satisfies $\frac{n}{n+N} < \frac{2s}{V_M}$. Similar arguments apply in other situations. ■

Note that if searchers *could* observe whether a firm's price was discounted or not, the equilibrium outcome would be that a searcher buys immediately if and only if the local price was discounted, and firms continue to set prices to reflect local demand conditions.[11] Thus, a simplifying feature of this particular framework is that equilibrium prices are not affected by policy towards misleading pricing.

Suppose the market initially operates in a regime where nothing except the current price is revealed to consumers. When does a firm have an incentive to reveal more details about its pricing policy? The firm's aim is simple: regardless of its current price state, it wishes to deter its local consumers from travelling to the remote firm. Suppose first that a firm can only make truthful claims about its prices. When the market state is $\{V_M, V_H\}$, a firm will announce that its price is discounted when $p = V_M$, as this will induce searchers to buy immediately (while otherwise they would have travelled to the other firm). Consumers are better off if they know when the local price is discounted, as this helps to refine their search strategy.

However, a firm has an incentive to mislead consumers, and falsely to claim its regular price is discounted. If firms are free to do so without penalty, savvy consumers will treat any claimed discount as cheap talk—they recognize that a firm will claim a price $p = V_M$ is discounted, regardless of whether the market state is $\{V_L, V_M\}$ or $\{V_M, V_H\}$—and so the outcome is as if consumers do not know whether or not the good is on sale. If instead consumers are more gullible, they believe a firm's false claims whenever such claims are possible. In this framework, this implies that when the market state is $\{V_L, V_M\}$ and a seller's price is $p = V_M$, the firm can claim its price is discounted (i.e., that the market state is $\{V_M, V_H\}$) and induce gullible consumers to buy immediately. (However, these consumers are not so gullible that they believe a firm's claim that its price $p = V_H$ was discounted.)

Expected expenditure from the searchers in the various regimes can be calculated as

[11] Condition (11) implies that a consumer will travel to the remote firm if the offered price $p = V_M$ when the consumers know the market state is $\{V_L, V_M\}$.

follows. In the regime where searchers do not know when a price is discounted, a searcher's expected outlay (including search costs where incurred) is[12]

$$\tfrac{3}{8}V_L + \tfrac{1}{2}V_M + \tfrac{1}{8}V_H + \tfrac{7}{8}s \ . \tag{12}$$

Likewise, when a searcher knows when a price is discounted, her expected outlay is

$$\tfrac{3}{8}V_L + \tfrac{1}{2}V_M + \tfrac{1}{8}V_H + \tfrac{1}{2}s \tag{13}$$

since she searches less often (although she makes exactly the same purchase decision). Finally, if the consumer is more gullible and always believes the price $p = V_M$ is discounted, her outlay is

$$\tfrac{1}{4}V_L + \tfrac{5}{8}V_M + \tfrac{1}{8}V_H + \tfrac{1}{4}s \ . \tag{14}$$

Here, relative to the other regimes, the consumer searches too little and ends up with a more expensive product on average.

In sum, in this stylized framework a policy which prevents firms from making misleading claims about discounts is good for consumers. With such a policy, a firm will always reveal when its price is discounted, and this enables consumers to improve their search strategy. Absent the policy, a firm will always claim its product is on sale, and consumers will be worse off: savvy consumers will disregard the permanent sale signs and search in ignorance of whether the local price is discounted or not; more gullible consumers will fall victim to the sale signs and too rarely search for a lower price. Industry profits are not affected by policy when consumers are savvy, as consumers make exactly the same purchase decisions in either regime. However, if consumers are more gullible, policy which prevents misleading price claims will reduce profits, as consumers are more likely to search for a better deal.

In general, the impact of policy on welfare depends on the underlying process of price determination, i.e., on whether profit margins are higher or lower when price is high or low. However, the impact is easy to understand in this framework where unit costs do not vary, since the prices paid by consumers are merely a transfer to firms and have no impact on welfare. Welfare is then inversely related to how much search occurs in the various regimes. By inspecting expressions (12)–(14), we see that welfare is highest when consumers are gullible and firms mislead them with false sales, for then search is rare. If instead consumers are savvy and disregard false sale signs, then policy to prevent misleading

[12]For instance, in this regime a consumer will pay the lowest price V_L when the market state is $\{V_L, V_M\}$ and at least one of the two firms has price $p = V_L$, which together occur with probability $\frac{3}{8}$. The consumer makes a costly trip with probability $\frac{7}{8}$, since she searches when the local price is not V_L and she makes *two* trips when the local price is V_M and the remote price is V_H.

sales signs reduces the intensity of search and so boosts welfare. In sum, while the impact of policy on consumers alone is clear-cut in this model, the impact on overall welfare is more complex and depends on the presumed gullibility of consumers.

We summarize this discussion as:

Proposition 2 *In the oligopoly search setting, when firms provide accurate information about when their price is discounted this benefits consumers relative to the situation where no such information is available. Consumers buy immediately when they see a discounted price. A policy which prevents firms from falsely claiming discounts will benefit consumers regardless of whether or not consumers believe false sales signs. The impact on welfare depends on whether consumers are gullible or savvy.*

4 Selling to Bargain-Loving Consumers

In our final model of discount pricing, we suppose that consumers intrinsically like the idea of getting a bargain. Thus, unlike models in sections 2 and 3, here we do not derive why it is that consumers care about receiving a discount, but simply take this as given. The model here, then, is a behavioural model with reference dependence. Unlike recent papers in industrial organization which focus on loss-aversion, we take the less familiar route of supposing consumers also enjoy a benefit if they pay a price below the reference price. In our model, the reference price is simply the average price offered by the firm.[13]

Suppose that a monopolist sells to a unit mass of consumers with constant marginal cost c, and chooses its price according to a mixed strategy with c.d.f. $G(p)$ which has expected value $\bar{p}$. (The firm offering a deterministic price as a special case of this framework.) Note that a given consumer is offered a single price, and cannot search for additional prices. To be concrete, we might imagine that the firm makes its price contingent on some arbitrary aspect of the consumer (e.g., location) which cannot easily be altered, and so pricing is not strictly random. Suppose a consumer's "raw" valuation for the item is v, which has smooth distribution function $F(v)$. If the consumer is given a "rip-off" price $p \geq \bar{p}$ then she buys if $v - \lambda_R(p - \bar{p}) \geq p$, where $\lambda_R \geq 0$ is a parameter which reflects her aversion

[13] An important ingredient of any model with reference dependence is how the reference point is determined. Broadly speaking, Heidhues and Kőszegi (2005) take the reference price to be the price a consumer expect to pay if she decides to buy, while Spiegler (2011b) takes the reference price to be the expected price *offered* by the seller (where that expected price is a random price draw from the firm, as might be generated by "word of mouth" for example). Puppe and Rosenkranz (2011) describe a model in which a manufacturer's non-binding "recommended retail price" acts as the reference price for consumers, while Zhou (2011) studies an oligopoly model in which consumers take the price of one "prominent" seller as their reference price when they evaluate other offers.

to paying above-average prices. If the consumer gets a bargain price $p \leq \bar{p}$ then she buys if $v + \lambda_B(\bar{p} - p) \geq p$, where $\lambda_B \geq 0$ is a parameter which reflects her enjoyment of the bargain.

Consider to start with the case where consumers are accurately informed about the firm's price policy (in particular, they know the average price $\bar{p}$, which, together with their own price, is what they care about). First, we show that it is always profitable for the monopolist to offer dispersed prices in this context, provided that consumers care more about getting a bargain than they do about avoiding a rip-off:

Lemma 4 *When consumers can observe the firms price policy, the firm prefers to offer dispersed prices than a uniform price when*

$$\lambda_B > \lambda_R \tag{15}$$

Proof. Let $p > c$ represent any profitable uniform price (not necessarily the most profitable uniform price). Suppose the firm deviates from this uniform price by offering two prices, $p_L = p - \varepsilon$ and $p_H = p + \varepsilon$ where $\varepsilon > 0$, where each price is offered to half the consumer population. (This modified strategy leaves the average price unchanged at p.) The firm's profit with this new strategy is

$$\pi(\varepsilon) \equiv \tfrac{1}{2}(p + \varepsilon - c)(1 - F(p + [1 + \lambda_R]\varepsilon)) + \tfrac{1}{2}(p - \varepsilon - c)(1 - F(p - [1 + \lambda_B]\varepsilon)) \ .$$

Differentiating this expression with respect to ε shows that

$$\pi'(0) = \tfrac{1}{2}(p - c)f(p)[\lambda_B - \lambda_R] > 0 \ ,$$

where $f(\cdot)$ is the density associated with $F(\cdot)$. Thus, starting from any profitable uniform price, profit is increased by implementing a mean-preserving spread in its prices. ■

The intuition for this result is clear. Relative to a uniform price strategy, adding a small amount of noise to prices reduces demand from those consumers offered above-average prices and boosts demand from those who get a bargain, and given (15) the latter effect dominates. We deduce that the firm has an incentive to offer at least two prices when consumers are more bargain-loving than loss-averse. Clearly, if only a fraction of consumers had these preferences (while the rest were "rational" and cared only about their own price), the firm would still have an incentive to pursue this dispersed pricing policy. If instead consumers were more loss-averse than bargain-loving, so $\lambda_B < \lambda_R$, then the firm has no (local) incentive to disperse its prices. In sum, the presence of bargain-loving consumers

gives the firm an incentive to offer distinct prices to otherwise identical consumers: in order to satisfy a "demand for bargains", the firm creates bargains by artificially dispersing its prices.

If we assume that the demand curve $1 - F$ is weakly concave, one can show that the firm will use *only* two prices in its optimal pricing policy. In order to derive this optimal policy, we suppose that the firm is restricted to offer prices which are sometimes accepted by consumers. (Or equivalently, that consumers ignore any price which is so high that demand at that price is zero when they calculate the average price.) Let $v_{\max}$ be the maximum valuation in the support of v. (Since the demand curve is concave, we know there is such a valuation.) Stated precisely, the firm is restricted to choose a price policy such that

$$p_{\max} + \lambda_R(p_{\max} - \bar{p}) \leq v_{\max} \ , \tag{16}$$

where $p_{\max}$ is the firm's maximum offered price and $\bar{p}$ is its expected offered price. This assumption rules out a strategy in which the firm offers arbitrarily high prices to a tiny fraction of consumers, which are not accepted, which would then make $\bar{p}$ arbitrarily large without significant cost to the firm.[14]

Lemma 5 *Suppose consumers have a preference for bargains in the sense that (15) holds and can observe the firm's price policy. If demand $1 - F(v)$ is weakly concave and the firm chooses prices which satisfy (16), the firm wishes to use exactly two prices in its pricing scheme.*

Proof. To avoid technicalities, suppose the firm offers a finite number of distinct prices (at least two in number), where price p_i is offered to a fraction $\alpha_i > 0$ of consumers and average price is $\bar{p} = \sum_i \alpha_i p_i$. Clearly, at least one price is strictly above the mean and one price is strictly below the mean.

Note first that it cannot be optimal for the firm to set any price below cost. (If some prices were below c, then profit is strictly increased by adjusting such prices to equal c: this adjustment increases $\bar{p}$ and so boosts demand from all consumers with $p_i \geq c$, and it clearly increases profit from these hitherto loss-making consumers.) So suppose that all prices satisfy $p_i \geq c$.

Next, we claim that the firm optimally offers only one price which is strictly above the mean. (The following argument is essentially an instance of Jensen's Inequality.) Suppose,

[14] A more satisfying solution to this problem would be for consumers to construct the "average price" in terms of the average *accepted* price among the consumer population instead of the firm's average offered price. However, this alternative approach is substantially more complex to solve.

to the contrary, there are at least two distinct prices, say p_1 and p_2, where $p_1 > p_2 > \bar{p}$. Suppose we reduce p_1 by $\varepsilon > 0$ and increase p_2 by $\frac{\alpha_1}{\alpha_2}\varepsilon$, where ε is small enough that both prices remain above $\bar{p}$ and that (16) continues to hold. By construction, the average price $\bar{p}$ is not affected by this change, and so the profits obtained from all other prices $p_i \notin \{p_1, p_2\}$ are unaffected. If we write $\pi(\varepsilon)$ for the firm's expected profits as a function of ε, then

$$\pi'(0) \stackrel{sign}{=} [F(p_1 + \lambda_R(p_1 - \bar{p})) - F(p_2 + \lambda_R(p_2 - \bar{p}))] \\ +(1 + \lambda_R)\left[(p_1 - c)f(p_1 + \lambda_R(p_1 - \bar{p})) - (p_2 - c)f(p_2 + \lambda_R(p_2 - \bar{p}))\right] .$$

This expression is strictly positive: the first term [.] is strictly positive since $F(.)$ is strictly increasing over this range, and the second term [.] is strictly positive from the assumption that $1 - F$ is weakly concave. We deduce that the original prices cannot be optimal, and so the firm chooses exactly one price above the average price in its optimal policy.

A similar argument shows that the firm's optimal policy also involves a single price which is weakly below the mean. ■

At least with concave demand, we deduce that the firm uses exactly two prices and so pursues a "high-low" price policy. It is then a simple matter to derive the firm's optimal price policy. If the firm offers the full price p_H with probability α and the discounted price $p_L < p_H$ with probability $1 - \alpha$, its profit is

$$(1-\alpha)(p_L-c)[1-F(p_L-\lambda_B\alpha(p_H-p_L))]+\alpha(p_H-c)[1-F(p_H+\lambda_R(1-\alpha)(p_H-p_L))] . \quad (17)$$

Consider the example where v is uniform on $[0, 1]$, $c = 0$ and $\lambda_R = 0$. Here, the most profitable uniform price is $p^* = \frac{1}{2}$. One can check from (17) that the optimal pricing strategy is

$$p_H = \frac{\sqrt{\lambda_B + 1} + 3}{8 - \lambda_B} \; ; \; p_L = \frac{p_H}{\sqrt{\lambda_B + 1}} \; ; \; \alpha = \frac{\sqrt{\lambda_B + 1} - 1}{\lambda_B} . \quad (18)$$

This policy satisfies $p_H > p^* = \frac{1}{2} > p_L$, so that the high price is above, and the low price is below, the optimal uniform price $p^* = \frac{1}{2}$. This solution requires λ_B to lie in the range $0 < \lambda_B < 3$ to satisfy (16). The policy converges to the optimal uniform price as λ_B becomes small. When $\lambda_B = 1$ the approximately optimal policy involves $p_L = 0.44$ and $p_H = 0.63$, and the full price is offered to 41% of consumers. Note that the average price here ($\bar{p} \approx 0.52$) is higher than it would be if the firm charged a uniform price (for instance, because consumers did not exhibit reference dependence, so $\lambda_B = 0$).[15] The

[15] Spiegler (2011a, section 9.1.2) shows that in a model where loss-aversion is the dominant force average price falls relative to the standard case.

firm's profit with this policy is about 0.26 and aggregate consumer surplus, taking their reference-dependent preferences at face value, is 0.15.

There are at least two ways to relax the strong assumption that consumers observe the firm's full pricing policy, and instead observe only the price they themselves are offered. First, savvy consumers could hold equilibrium beliefs about the average price; second, consumers might be more gullible and believe the firm's claims about its average price.[16]

Consider first the situation where consumers hold equilibrium beliefs about the firm's entire pricing strategy, even though they observe only their own price. That is to say, from a consumer's viewpoint, the firm's prices to other consumers are "secret". If all consumers believe the average price is P, the firm's expected profit when it offers price p to a given consumer is $(p-c)(1-F(p-\lambda_B(P-p)))$ if $p \leq P$ and $(p-c)(1-F(p+\lambda_R(p-P)))$ otherwise.[17] Thus, when (15) holds the firm faces a demand curve with an "inward" kink at the reference price P. In this case we have the following result.[18]

Lemma 6 *Suppose consumers observe only their own price, and that the demand curve* $1-F(\cdot)$ *is logconcave.*[19] *If (15) holds then (i) there is no equilibrium in which the firm offers a uniform price, and (ii) there exists an equilibrium in which the firm offers exactly two prices,* p_L *and* p_H*, where both of the these prices are below the most profitable uniform price* p^**.*

Proof. (i) If to the contrary P is an equilibrium uniform price, anticipated by consumers, the firm cannot make greater profit by choosing $p < P$, so that

$$1 - F(P) - (1+\lambda_B)(P-c)f(P) \geq 0 \ ,$$

and neither can the firm make greater profit by choosing $p > P$, so that

$$1 - F(P) - (1+\lambda_R)(P-c)f(P) \leq 0 \ .$$

These two inequalities are inconsistent if (15) holds.

[16] In this paper we assume that the firm either makes all its prices public or none. An interesting variant is to suppose that the firm can selectively reveal is price policy to consumers, in which case it might reveal the average price to those consumers who get a bargain, but keep those who pay a high price in the dark.

[17] Here, we assume consumers have "passive beliefs" about the average price, and the price p a consumer is offered does not alter her anticipated P.

[18] In formal terms, this result resembles the analysis in Zhou (2011). Like us, he finds that a seller faces demand with an inward kink and chooses prices according to a mixed strategy with exactly two prices; in his case, the prominent seller uses "sales" to influence a loss-averse consumer's reference point when she evaluates the rival offer, while our firm uses "sales" to satisfy a consumer's demand for bargains.

[19] If $1-F$ is weakly concave it is also logconcave.

(ii) We construct the "high-low" equilibrium as follows. Let consumers anticipate the average price P. If the firm chooses a price strictly above P, this price p_H must (locally) maximize $(p-c)(1-F(p+\lambda_R(p-P)))$, and when demand is logconcave there is at most one such price, which is determined for given P by the first-order condition

$$p_H = c + \frac{1-F(p_H+\lambda_R(p_H-P))}{(1+\lambda_R)f(p_H+\lambda_R(p_H-P))} \ . \tag{19}$$

Likewise, if the firm chooses a bargain price below P, this price p_L must maximize $(p-c)(1-F(p-\lambda_B(P-p)))$, which is uniquely determined for given P by the first-order condition

$$p_L = c + \frac{1-F(p_L-\lambda(P-p_L))}{(1+\lambda_B)f(p_L-\lambda(P-p_L))} \ . \tag{20}$$

The firm must be indifferent between choosing the two prices p_L and p_H, so that

$$(p_L-c)(1-F(p_L-\lambda_B(P-p_L))) = (p_H-c)(1-F(p_H+\lambda_R(p_H-P)) \ . \tag{21}$$

Finally, in equilibrium consumer expectations of the average price are fulfilled, so that

$$P = \alpha p_H + (1-\alpha)\, p_L \tag{22}$$

where α is the fraction of consumers who pay p_H. The four tariff parameters p_L, p_H, P and α then solve the four equations (19)–(22).

To see that a solution to these four equations exists, argue as follows. First note that if we can find p_L, p_H and P satisfying (19)–(21) such that $p_L < P < p_H$, then we can find an $0 < \alpha < 1$ which satisfies (22). Therefore, we look for p_L, p_H and P satisfying (19)–(21) such that $p_L < P < p_H$. Since $1-F(\cdot)$ is logconcave, we can check that p_H in (19) is above P if and only if P is sufficiently small, and the threshold P which makes the firm choose $p_H = P$ in (19) is

$$P_H = c + \frac{1}{1+\lambda_R} \cdot \frac{1-F(P_H)}{f(P_H)} \ .$$

Likewise, from (20) we can see that p_L is below P when P is sufficiently large, and the threshold P which makes the firm choose $p_L = P$ in (20) is

$$P_L = c + \frac{1}{1+\lambda_B} \cdot \frac{1-F(P_L)}{f(P_L)} \ .$$

Given the logconcavity of $1-F$ and assumption (15), it follows that $P_L < P_H$. Thus, for any P in the range $P_L < P < P_H$, the firm's high price in (19) is above P and the firm's discounted price in (20) is below P. Note that both P_L and P_H are below p^*, the optimal uniform price.

It remains to show that we can find P in the range $P_L < P < P_H$ such that (21) holds. Consider the lower boundary $P = P_L$. By construction, when $P = P_L$ then $p_L = P_L$ in (20) in which case the firm's profit when it chooses $p = p_L$ is $(P_L - c)(1 - F(P_L))$. But when $P = P_L$, the firm's profit when it chooses p_H in (19) is strictly higher than this, since the firm could have chosen $p_H = P_L$ which yields the same profit $(P_L - c)(1 - F(P_L))$. Thus, when $P = P_L$ the firm makes strictly greater profits by choosing p_H in 19) than it does by choosing p_L in (20). A similar argument establishes that when $P = P_H$, the firm does strictly better by choosing the lower price p_L in (20) than by choosing p_H in (19). By continuity, there exists at least one P in the range $P_L < P < P_H$ where the firm is indifferent between choosing p_L in (20) and p_H in (19). This completes the proof. ■

In the same example where v is uniform on $[0, 1]$, $c = 0$ and $\lambda_R = 0$, the equilibrium pricing policy in the regime where consumers observe only their own price can be shown from expressions (19)–(22) to be

$$p_H = p^* = \frac{1}{2} \; ; \; p_L = \frac{p_H}{\sqrt{\lambda_B + 1}} \; ; \; \alpha = P = \frac{\sqrt{\lambda_B + 1} - 1}{\lambda_B} \; . \tag{23}$$

Note that the high price in this example is equal to the optimal uniform price, and from (19) this is true whenever $\lambda_R = 0$ so that consumers do not care when they pay an above-average price. When $\lambda_B = 1$, the firm's profit as a function of its price p offered to any particular consumer, given that the consumer believes average price is $P = \sqrt{2} - 1$, looks as shown on Figure 1. This figure illustrates the bimodal nature of profit with bargain-loving consumers, and the equilibrium is constructed so that the height of the two peaks coincides.

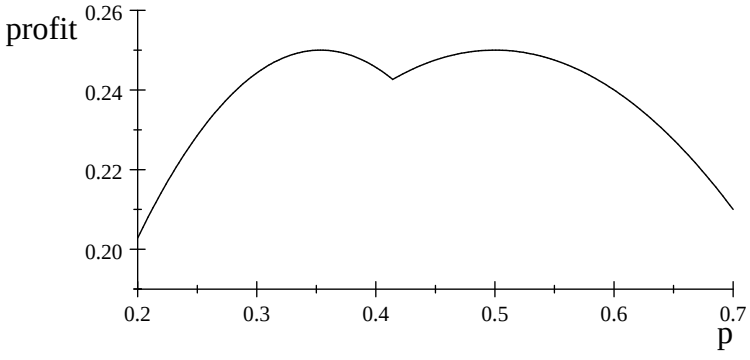


Figure 1: Monopolist's profit as function of p

This price policy in (23) is qualitatively the same as in the case in (18) where a consumer can observe the firm's prices for all consumers; in particular the percentage discount p_L/p_H

is the same and the likelihood of getting a bargain is the same. However, prices are now shifted downwards. Of course, quite generally, the firm's profits here are lower compared to when consumers see the full range of prices, since the firm could choose the pricing policy seen with secret deals as its policy when its prices are public. In this linear demand example, aggregate consumer surplus is now higher, at about 0.2, and total welfare is higher when the firm's prices are privately observed. Intuitively, when the firm makes secret deals with each consumer, the firm has a greater incentive to undercut the average price since other consumers do not observe, and cannot react to, the price cut.[20]

Suppose that the firm is able to make false claims about its average price. If consumers are savvy, they foresee that the firm has an incentive to exaggerate its average price to boost its demand from bargain-loving consumers, and so consumers discount its claims and behave as if they cannot observe the average price. In such a situation, a policy which enables the firm credibly to reveal its average price will help the firm and, at least in the linear demand example, *harm* consumers. If policy forces the firm to publish accurate information about its prices and the proportion of prices which are discounted, then any price-cut targeted at particular individuals reduces demand from other consumers, and so blunts the firm's incentive to discount.[21]

On the other hand, if consumers are more gullible and believe its claims, the firm's profits are increased when it is able to make misleading claims. It can then obtain the benefit of boosting demand from perceived "bargains" without the cost of sometimes having to set inefficiently high prices. It would like to claim average price was as high as possible, so that it could then set high actual prices without cutting demand.[22]

Summarizing our discussion of this model, we have:

Proposition 3 *(a) Suppose consumers have an intrinsic preference for bargains. Then the monopolist will offer distinct prices to identical consumers. If demand* $1 - F(p)$ *is weakly concave, the firm will adopt a "high-low" pricing strategy and offer exactly two prices to the population of consumers. The firm's profit is higher when consumers can observe its average price compared to when they have no information about its average price.*

[20] The effect is analogous to the "secret deals" problem in vertical contracting, discussed in Rey and Tirole (2007), in which an upstream manufacturer who sells to two competing retailers has an opportunistic incentive to boost supply to a retailer when the other does not observe the deal.

[21] Again, this is similar to the impact of policy on the secret deals problem in vertical contacting, where a requirement to make the supplier's deal to one retailer observed by another will boost supplier profits and harm final consumers.

[22] In this case, the welfare impact of a policy banning false discounts is more complicated, and depends on how one views a consumer's utility from getting a "false bargain".

(b) Suppose demand is linear. A policy which prevents the firm from making false claims about its average price helps the firm and harms consumers and welfare if consumers are savvy and foresee the firm will exaggerate its average price. The same policy will harm the firm if consumers are gullible and believe its claims about average price.

5 Conclusion

This paper has explored some economic effects of discount pricing. We suggest two reasons why a discounted price—as opposed to a merely low price—may make a rational consumer more willing to buy. First, the information that the product was initially sold at a high price may indicate the product is high quality. Second, a discounted price can indicate that the product is an unusual bargain, and that there is little point searching for alternative, lower prices. We also discuss discount pricing with behavioural consumers. If consumers have an intrinsic preference for bargains, a seller has an incentive to offer different prices to identical consumers, so that a proportion of its consumers will enjoy a bargain. Information about discounts in this case assures consumers how good their deal is relative to the average, which boosts their willingness to purchase.

Because of their incentive to mislead customers, in some—but not all—of the situations we discuss, there is a potential role for policy to prevent sellers advertising false discounts. In all models, if consumers are gullible and believe—rather than merely ignore—a firm's false claims, such a policy will help consumers and harm the firm. In most cases, the overall impact on welfare of a policy which combats false discounting is positive.[23] If consumers are savvier, matters are more nuanced. In our model where the initial price serves to signal the choice of high quality, a ban on misleading claims will actually benefit the firm, as it makes it easier to signal its quality. In the model with oligopoly search, such a policy benefits consumers as they then learn when an offered price is a discounted price and can reduce their search effort. Finally, in our model of bargain-lovers, when consumers are savvy a ban on misleading price claims will help the firm but *harm* consumers. Policy which helps the firm make public its pricing policy overcomes its "secret deals" problem, to the detriment of consumers.

In any case, the potential benefit from regulatory policy can be realized only if it is effectively enforced. Indeed, weakly enforced policy may be worse than no policy: it may make consumers gullible and act on a firm's false discounts, and it may harm honest sellers

[23] The exception is the model of oligopoly search in section 3, where permanent sale signs induce gullible consumers to buy more often from their local seller, which reduces search costs.

who follow the letter of policy. As discussed by Muris (1991) and Rubin (2008), it is hard to enforce, or perhaps even coherently to formulate, policy towards misleading pricing. A basic problem is how to determine how few sales need to occur at the full price, or for how short a time the full price is available, for a sales campaign stating "was $200, now $100" to be classified as misleading. Sellers have a strong motive to make their customers feel they are getting a special deal, and they have myriad ways to achieve this. It is unrealistic and undesirable to suppose that regulation can address all forms of false discounting without unduly restricting a seller's marketing abilities, and regulators should focus only on flagrant examples of deception.

References

Bagwell, Kyle and Michael Riordan (1991), "High and Declining Prices Signal Product Quality", *American Economic Review* 81(1), 224-239.

Bordalo, Pedro, Nicola Gennaioli and Andrei Shleifer (2012), "Salience and Consumer Choice", NBER working paper 17947.

Cialdini, Robert (2001), *Influence: Science and Practice*, Allyn and Bacon (4th Edition).

Heidhues, Paul and Botond Kőszegi (2005), "The Impact of Consumer Loss Aversion on Pricing", CEPR discussion paper no. 4849.

Jahedi, Salar (2011), "A Taste for Bargains", mimeo, University of Arkansas.

Lazear, Edward (1986), "Retail Pricing and Clearance Sales", *American Economic Review* 76(1), 14–32.

Muris, Timothy (1991), "Economics and Consumer Protection", *Antitrust Law Journal* 60, 103-121.

Puppe, Clemenz and Stephanie Rosenkranz (2011), "Why Suggest non-binding Retail Prices?", *Economica* 78, 371-329.

Rey, Patrick and Jean Tirole (2007), "A Primer on Foreclosure", chapter 33 in *Handbook of Industrial Organization vol. 3*, edited by Armstrong and Porter, Amsterdam: North-Holland.

Rubin, Paul (2008), "Regulation of Information and Advertising", *Competition Policy International* 4(1), 169–192.

Spiegler, Ran (2011a), *Bounded Rationality and Industrial Organization*, Oxford University Press.

Spiegler, Ran (2011b), "Monopoly Pricing when Consumers are Antagonized by Unexpected Price Increases: A 'Cover Version' of the Heidhues-Koszegi-Rabin Model", *Economic Theory* (forthcoming).

Thaler, Richard (1985), "Mental Accounting and Consumer Choice", *Marketing Science* 4(3), 199-214.

Tirole, Jean (1988), *The Theory of Industrial Organization*, MIT Press.

Urbany, Joel, William Bearden and Dan Weilbaker (1988), "The Effect of Plausible and Exaggerated Reference Prices on Consumer Perceptions and Price Search", *Journal of Consumer Research* 15(1), 95-110.

Zhou, Jidong (2011), "Reference Dependence and Market Competition", *Journal of Economics and Management Strategy* 20(4), 1073-1097.

Cellular Service Demand: Biased Beliefs, Learning, and Bill Shock*

Michael D. Grubb† and Matthew Osborne‡

February 27, 2012

Abstract

By April 2013, the FCC's recent bill-shock agreement with cellular carriers requires consumers be notified when exceeding usage allowances. Will the agreement help or hurt consumers? To answer this question, we estimate a model of consumer plan choice, usage, and learning using a panel of cellular bills. Our model predicts that the agreement will lower average consumer welfare by $2 per year because firms will respond by raising monthly fees. Our approach is based on novel evidence that consumers are inattentive to past usage (meaning that bill-shock alerts are informative) and advances structural modeling of demand in situations where multi-part tariffs induce marginal-price uncertainty. Additionally, our model estimates show that an average consumer underestimates both the mean and variance of future calling. These biases cost consumers $42 per year at existing prices. Moreover, absent bias, the bill-shock agreement would have little to no effect.

*A previous version circulated under the title "Cellular Service Demand: Tariff Choice, Usage Uncertainty, Biased Beliefs, and Learning". We thank Parker Sheppard and Mengjie Ding for research assistance and Katja Seim, Panle Jia, Eugenio Miravete, Catherine Tucker, Greg Lewis, Chris Knittel, Ron Goettler, and S. Sriram for careful reading and feedback on early drafts. We also thank Ted O'Donoghue and seminar audiences at Duke, Cornell, Chicago, and Rochester for useful feedback. The views expressed herein are those of the authors and not necessarily those of the Bureau of Economic Analysis or the U.S. Department of Commerce. Work on this paper began when Osborne worked at the Department of Justice.

†Massachusetts Institute of Technology, Sloan School of Management. mgrubb@mit.edu.

‡Bureau of Economic Analysis. Matthew.Osborne@bea.gov.

1 Introduction

Cellular phone companies frequently offer consumers contracts with included allowances of voice minutes, text messages, and data usage that are followed by overage charges for higher usage. Consumers are often unaware that they are incurring overage charges during the month, which leads to *bill shock* at the end of the month. On October 17^{th}, 2011 President Barack Obama declared:

> Far too many Americans know what its like to open up their cell-phone bill and be shocked by hundreds or even thousands of dollars in unexpected fees and charges. But we can put an end to that with a simple step: an alert warning consumers that they're about to hit their limit before fees and charges add up (CTIA - The Wireless Association 2011a).

President Obama made this statement at the announcement of a new bill-shock agreement between the FCC and cellular carriers. By April 2013, this agreement commits cellular service providers to inform consumers when they approach and exceed their included voice, text, and data allowances (CTIA - The Wireless Association 2011a). Prior to the agreement, the FCC had proposed a similar regulation which was strongly supported by consumer groups but opposed by the industry (Deloney, Sherry, Grant, Desai, Riley, Wood, Breyault, Gonzalez and Lennett 2011, Altschul, Guttman-McCabe and Josef 2011).[1]

Will the new bill-shock agreement help or hurt consumers? If carriers held their prices fixed after implementing the agreement then it would weakly help consumers. Such prices-fixed logic likely lies behind consumer groups' strong advocacy for bill-shock alerts. However, the bill-shock agreement could hurt consumers once endogenous price changes are taken into account. Moreover, complementary theoretical work by Grubb (2011) shows that the answer is theoretically ambiguous. Therefore, to address this question, we develop and estimate a dynamic model of plan choice and usage that makes use of detailed cellular phone data. Given our parameter estimates, counterfactual simulations show that the net effect of the bill-shock agreement and endogenous prices changes is an overall annual reduction in consumer welfare of $2 per consumer.

En route to making our prediction about the bill-shock agreement's effect on consumer welfare we make two additional contributions. First, we provide new evidence on how consumers make consumption choices under marginal-price uncertainty and estimate a tractable model incorporat-

[1]The wirelesss industry trade group, C.T.I.A. - The Wireless Association, argued that proposed bill-shock regulation "violates carriers' First Amendment protections. . . . against government compelled speech" (Altschul et al. 2011).

ing such realistic behavior. In particular, we find that consumers are inattentive to their remaining minute balance. Given such inattention, we assume that consumers optimally respond to exogenously arising calling opportunities by choosing a calling threshold and making only those calls more valuable than the threshold. Unlike standard models, this approach allows for consumers to endogenously adjust their calling behavior in response to bill-shock alerts in our counterfactual simulations. (Attentive consumers would never find new information in a bill-shock alert.) Second, we relax the standard rational expectations assumption and infer consumers' beliefs about their future calling opportunities from plan choices. By comparing these beliefs to actual usage at the population level, systematic differences identify consumer biases such as overconfidence. Identifying consumer biases is important for our endogenous-price counterfactual simulations because firm pricing decisions are strongly influenced by overconfidence and other biases (Grubb 2009).

Our primary data were obtained from a major US university that acted as a reseller for a national cellular phone carrier, and covers all student accounts managed by the university from 2002 to 2004. We begin by documenting five stylized facts in our data that shape our modelling approach. First, a sharp increase in calling when free off-peak calling begins shows that consumers' usage choices are price sensitive. Second, absence of bunching at tariff kink points and other evidence show that consumers are uncertain about the ex post marginal price when making calling choices. Third, novel evidence from call-level data suggests consumers are inattentive to their remaining balance of minutes. Fourth, consumers are uncertain about their own average taste for usage when first choosing a calling plan, which leads to frequent ex post plan choice mistakes. However, consumers learn about their own tastes over time and switch plans in response. Finally, consumers make ex ante mistakes that are predictable given information held by a carrier.

The first three stylized facts suggest that the arrival of a bill-shock alert will be informative and cause a consumer to reduce calling. The second stylized fact, marginal-price uncertainty, naturally arises whenever consumers make a series of small purchase choices that are aggregated and billed under a multipart tariff, as in cellular phone service, electricity, and health care. Addressing such marginal-price uncertainty represents a challenge for the literature which has typically side-stepped the issue by assuming that consumers can perfectly predict their future usage (Cardon and Hendel 2001, Reiss and White 2005, Lambrecht, Seim and Skiera 2007), or that consumers believe they can perfectly predict their usage up to an implementation error which they ignore (Iyengar, Ansari and Gupta 2007). (Notable exceptions are Yao, Mela, Chiang and Chen (2011) and Jiang (2011).) By recognizing that consumers are inattentive, our modeling approach incorporates marginal-price uncertainty realistically and tractably and allows consumers to endogenously respond to bill-shock alerts. Our consumers behave optimally given their inattention, by choosing a calling threshold each

month (related to expected marginal price) and accepting only calls valued above the threshold. This approach has been proposed in earlier work (Saez 2002, Borenstein 2009), but has not been implemented in a structural model.[2] An advantage of our structural approach is that we can estimate the consumer beliefs required to calculate calling thresholds.

To account for the last two stylized facts concerning plan choice, we model consumer beliefs and learning. We call a consumer's average taste for calling his true type. A consumer's plan choices are determined not by his true type but by his beliefs about his true type. We assume that each consumer's prior consists of a point estimate of her own true type and a level of perceived uncertainty about this point estimate. We assume that consumers are Bayesian learners, following Erdem and Keane (1996), Ackerberg (2003), Crawford and Shum (2005), and Goettler and Clay (2011) and therefore learn their true types in the long run. At the same time, to account for the predictable nature of plan choice mistakes in the short run, we allow consumers' initial beliefs to be biased.

Our data are informative both about consumers' actual average tastes for cellular phone usage and about their prior beliefs about their own tastes. Consumers' usage choices identify the distribution of consumers' true types, while consumers' initial plan choices and subsequent switching decisions identify beliefs. The joint distribution of beliefs and true types determines whether beliefs are biased in the population. For instance, suppose that we consider the subset of consumers that all share a particular prior belief about their own types. A common assumption (often labeled rational expectations) is that this belief coincides with the distribution of true types within this subset of the population. We relax this assumption, separately identify both beliefs and the distribution of true types conditional on beliefs, and then compare the two distributions. We label differences between these distributions as biases.[3] Moreover, we allow consumers to over- or under-estimate the monthly volatility in their tastes.

We identify two substantial biases causing predictable mistakes. The first we label *overconfidence*, which arises when a consumer underestimates her own uncertainty surrounding her point estimate of her true type. We find that consumers underestimate their own uncertainty about their true type by 84%. Overconfident consumers initially choose plans that are too risky. Moreover,

[2]In the context of electricity demand, Borenstein (2009) independently proposes that consumers choose *behavioral rules*, such as setting the thermostat, similar to our calling threshold. Borenstein (2009) uses the behavioral rule assumption to motivate using expected marginal price rather than realized marginal price in reduced form estimates of electricity price elasticities. Saez (2002) also suggests a very similar model for labor choice by income tax filers.

[3]An alternate interpretation is that unmeasurable prior beliefs were unbiased at some previous time, but are now measurably and systematically different from reality at the population level (although consistent with rational expectations) due to the arrival of a correlated shock or signal at the population level. The distinction is pedantic as it does not matter for optimal firm pricing, consumer welfare, policy counter-factuals or other issues of interest.

they place too much weight on their prior point estimates when updating beliefs and will be slow to learn and switch plans based on experience.

The second bias that we focus on is *volatility bias*, which arises when consumers underestimate the monthly volatility in their tastes for usage.[4] We estimate that consumers underestimate the volatility in their taste for usage by 10%. Similar to overconfidence, volatility bias causes consumers to underestimate the uncertainty in their usage predictions when making plan choices, and choose plans that are too risky. However, volatility bias has the opposite effect of overconfidence on the rate of learning: volatility bias causes consumers to underweight their priors relative to past usage when updating their beliefs about their average tastes for usage. This leads to faster learning and more frequent plan switching. Thus the rate of plan switching allows us to separate the two biases.[5] Because we find that overconfidence is stronger than volatility bias, consumers overweight their prior beliefs relative to new information and learn and switch plans relatively slowly. Thus initial plan choice mistakes are especially costly. Holding observed prices constant, we find that overconfidence and volatility bias jointly reduce annual consumer welfare by $34 per student.

There are other biases in beliefs which could result in consumer behavior that is similar to that caused by overconfidence and volatility bias. To ensure we do not misattribute other errors as overconfidence or volatility bias, we estimate a flexible distribution of initial beliefs which captures (at least) two other potential sources of bias. We are able to separately identify these biases due to the rich choice set of plans in our data that importantly include both three-part tariffs and a two-part tariff. Holding observed prices constant, these biases reduce annual consumer welfare by an additional $8 per student, for a total annual cost of all biases of $42 per student.

Turning back to the recent FCC agreement, we conduct a counterfactual simulation where we allow firms to adjust prices in response to bill-shock alerts. To do so, we add additional supply side structure to our model and add a parameter λ measuring the amount of differentiation across firms. This firm differentiation parameter λ is omitted from our estimated demand model because our demand data are from a single carrier and do not identify λ. To complete our endogenous price counterfactual simulations, we therefore first calibrate the firm differentiation parameter λ conditional on our demand estimates using observed prices. (We use EconOne data on the prices of all cellular-phone plans offered during 2002-2004 in the vicinity of the university that provided our primary data.) We find that firms respond to bill-shock regulation by raising fixed fees, reducing

[4]Overconfidence could more broadly be interpreted to include volatility bias, however we seek to draw a distinction between two different biases and define overconfidence more narrowly to do so.

[5]Our model includes a price consideration parameter that plays a similar role to a switching cost. This is separately identified from the learning rate by the rate at which consumers fail to switch away from strictly dominated plans.

included minute allowances, and reducing overage rates on three-part tariffs. By doing so, firms maintain annual profits close to unregulated levels (falling by just $0.20 per person).[6] This means that consumers are approximately residual claimants on total welfare, which falls by $2.21 per person, and hence annual consumer welfare drops by $2.01 per person. The social welfare loss results from consumers' reduced calling. Absent consumer biases, we find that firms offer two-part tariffs but not three-part tariffs, which means that bill-shock regulation has no effect.

Section 2 discusses related literature. Section 3 describes our data and documents five stylized facts that shape our modeling approach. Sections 4 and 5 describe our model and explain identification. Sections 6, 7, and 9 discuss estimation, present results and conclude. Additional details are in the Online Appendix available at \url{www.mit.edu/~mgrubb/GrubbOsborneAppendix.pdf}.

2 Related Literature

Complementary work by Jiang (2011) also evaluates the recent bill-shock agreement via counterfactual simulation, predicting a $370 million welfare improvement. In contrast to our own approach, Jiang (2011) imposes rational expectations rather than estimating consumer beliefs and has cross-sectional data so cannot address learning. Finally, Jiang's (2011) bill-shock counterfactual corresponds to removing a taste shock from the model. In contrast, a strength of our approach is that consumers endogenously change calling behavior in response to information in bill-shock alerts. (A strength of Jiang's (2011) data is that they are nationally representative and cover all carriers.)

Related work provides evidence that individual labor choices (Liebman and Zeckhauser 2004) and electricity consumption choices (Ito 2010) respond to average prices rather than marginal prices. This is not surprising because electricity tariffs and the income tax code are both very complex and often not well understood by consumers. A typical consumer may not realize electricity pricing is nonlinear, in which case average price is a good estimate of marginal price. However, this model is not appealing in the context of cellular service because consumers are fully aware that contracts include an allowance of 'free' minutes.

A significant body of experimental evidence shows that individuals are overconfident about the precision of their own predictions when making difficult forecasts (e.g. Lichtenstein, Fischhoff and Phillips (1982)). In other words, individuals tend to set overly narrow confidence intervals relative to their own confidence levels. A typical psychology study might pose the following question to a group of subjects: "What is the shortest distance between England and Australia?" Subjects would then be asked to give a set of confidence intervals centered on the median. A typical finding

[6] Annual profits fall by $3.69 per consumer for any single firm that independently chooses to offer bill-shock alerts.

is that the true answer lies outside a subject's 98% confidence interval about 30% to 40% of the time. Consumers who exhibit volatility bias underestimate the extent to which their tastes will change over time. This is closely related to projection bias, a prevalent behavior that has been documented in a variety of experiments, surveys, and field studies (Loewenstein, O'Donoghue and Rabin 2003, Conlin, O'Donoghue and Vogelsang 2007). Via mean biases, we allow for overestimation of demand, which is one of the causes of flat-rate bias documented by Lambrecht and Skiera (2006) in internet service choice.

A small number of empirical papers relax rational expectations for consumer beliefs and estimate mean biases (Crawford and Shum 2005, Goettler and Clay 2011). Most similar to our work is Goettler and Clay (2011), which estimates mean biases. Goettler and Clay (2011) cannot identify higher moments of beliefs because the choice set in online grocery-delivery service is limited to two-part tariffs. In contrast, the rich tariff choice-set in our setting enables us to measure (rather than assume away) volatility bias and overconfidence in addition to mean biases.

To identify beliefs from plan choices, we assume consumers are risk neutral.[7] In contrast, related work on health insurance markets often does the reverse and imposes rational expectations to identify risk preferences from plan choices (Cardon and Hendel 2001, Handel 2011, Einav, Finkelstein, Pascu and Cullen Forthcoming). Following a third approach, Ascarza, Lambrecht and Vilcassim (2012) impose rational expectations and risk neutrality but estimate preferences for cellular phone usage that depend directly on whether contracts are two or three-part tariffs.

Our results are consistent with a related sequence of papers about Kentucky's 1986 local telephone tariff experiment (Miravete 2002, Miravete 2003, Miravete 2005, Narayanan, Chintagunta and Miravete 2007, Miravete and Palacios-Huerta 2011). First, although the standard model of consumer choice does well at explaining behavior in the Kentucky experiment, our estimates of negative aggregate mean bias and positive conditional mean bias are consistent with evidence in Miravete (2003) which documents that on average all consumers who chose a small metered plan would have saved money on a larger flat rate plan.[8] Second, as in the Kentucky experiment we find that most consumers (55 to 71 percent) initially choose the tariff that turns out to be optimal ex post. Moreover, consumers switch plans and most switches appear to be in the right direction to lower bills (Section 3.2).

Our counterfactual simulations with endogenous prices relate to the literatures with standard

[7] If consumers are risk averse then our estimates of overconfidence and volatility bias are lower bounds on bias.

[8] Interestingly, in Miravete (2003) the bias that can be inferred from elicited expectations differs from that inferred from choices. Consumers were not offered three-part tariffs in the Kentucky experiment so their choices do not shed light on overconfidence or volatility bias.

consumers on monopoly sequential-screening (surveyed by Rochet and Stole ((2003), Section 8), including Baron and Besanko (1984), Riordan and Sappington (1987), Miravete (1996), Courty and Li (2000), Miravete (2005), and Grubb (2009)) and competitive static-screening (surveyed by Stole (2007), including Armstrong and Vickers (2001) and Rochet and Stole (2002)). Moreover, it is related to the growing literature on optimal contracting with non-standard consumers (for which Spiegler (2011) provides a good guide). Of particular relevance are DellaVigna and Malmendier (2004), Uthemann (2005), Eliaz and Spiegler (2006), Eliaz and Spiegler (2008), Grubb (2009), Herweg and Mierendorff (Forthcoming), and Grubb (2011).

Finally our paper is about the cellular phone industry, about which there is a small literature. Beyond work already mentioned, other work on the cellular phone industry examines risk while driving (Bhargava and Pathania 2011), carrier switching costs (Kim 2006), the effect of entry on pricing (Seim and Viard 2010, Miravete and Röller 2004), the effect of number portability regulation on competition (Park 2009), the role of multi-market contact in competition (Busse 2000), and demand (Iyengar, Jedidi and Kohli 2008, Huang 2008).

3 Background: Data and Evidence for Stylized Facts

3.1 Data

Our primary data are a panel of individual monthly billing records for all student enrollees in cellular-phone plans offered by a national cellular carrier in conjunction with a major university from February 2002 to June 2005. During this period, cellular phones were a relatively new product in the US, having 49% penetration in 2002 compared to 98% in 2010.[9] This data set includes both monthly bill summaries and detailed call-level information for each subscriber.[10] We also acquired EconOne data on the prices and characteristics of all cellular-phone plans offered at the same dates in the vicinity of the university. The price menu offered to students differed from that offered by the carrier directly to the public: university plans included a two-part tariff, a limited three-month contractual commitment, different monthly promotions of *bonus* minutes, and a $5 per month surcharge on top of carrier charges to cover the university's administrative costs.

[9]This feature makes our data ideal for studying consumer beliefs about new products. Penetration rates are calculated as estimated total connections (CTIA - The Wireless Association 2011b) divided by total population (U.S. Census Bureau 2011).

[10]Students recieved monthly phone bills, mailed by default to their campus residence. The sample of students is undoubtedly different than the entire cellular-phone-service customer-base. However, a pricing manager from one of the top US cellular phone service providers made the unsolicited comment that the empirical patterns of usage, overages, and ex post "mistakes" documented in Grubb (2009) using the same data were highly consistent with their own internal analysis of much larger and representative customer samples.

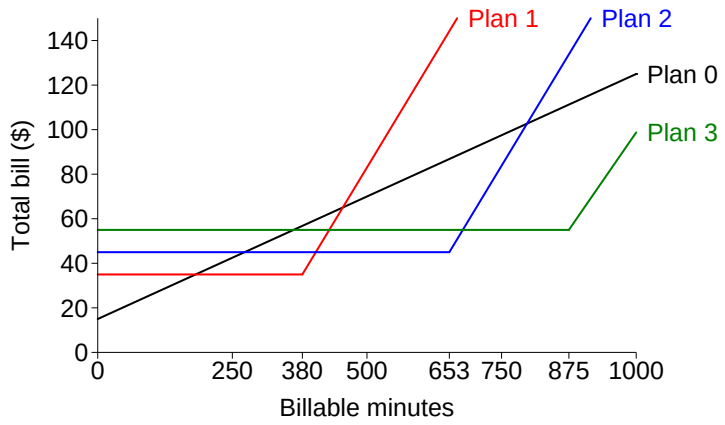
The bulk of our work makes use of the monthly billing data. We exclude individuals who are left censored (those who are existing subscribers at the start of the panel). For most analysis, including our structural estimation, we also restrict attention to the period August 2002 to July 2004. (This is the period for which we can reliably infer university prices from billing data. See Appendix A). We focus on customer choice between four *popular* local plans, that account for 89% of bills in our data. We group the remaining price plans (including national and free long distance plans) with the outside option, and hence drop the 11% of bills with unpopular price plans.[11] Finally, rate plan codes are frequently miscoded as a default value on a customers initial bill, in which case we remove the first bill. Our final data set contains 1366 subscribers and 16,283 month-subscriber observations. Note that for much of our analysis, we also exclude pro-rated bills during months of partial service, or customer switching between plans (however, pro-rated bills are included in the sample we use to estimate the structural model).

Figure 1 shows the four popular plans, which we label as plans 0 through 3. Plan 0 is a two-part tariff that charges $14.99 per month and 11 cents per minute. Plans 1-3 are three-part tariffs that charge monthly fees (M_j) of 34.99, 44.99, and 54.99 respectively, include an allowance (Q_j) of 280 to 1060 free peak-minutes, and charge an overage rate (p_j) of 35 to 45 cents per additional peak minute. We say that one plan is larger than another if it coincides with the lower envelope of the tariff menu at a higher interval of usage. Plans are numbered in order of size, smallest to largest. We say that one plan is riskier than another if it yields higher expected bills for sufficiently high usage uncertainty. Plan 0 is the safest plan, plan 1 is the riskiest, and plans 1-3 are numbered in order of decreasing risk.

All four plans include surcharges of 66 to 99 cents per minute for roaming outside a subscriber's tri-state area and 20 cents per minute for long distance. Plans 1-3 always offer free off-peak calling but Plan 0 does so only prior to fall 2003. Plan 0 includes free in-network calling, while plans 1-3 do not with the exception of plan 2 in 2004. Once a customer chooses a plan, the plan terms remain fixed for that customer, regardless of any future promotions or discounts, until they switch plans or terminate service. However, the terms of any given plan, such as the included allowances and overage rates for plans 1-3, vary according to the date a customer chooses the plan.

Shares of plans 0-3 are 44, 28, 15, and 2 percent of bills, respectively. Plan prices are shown for Spring 2003 in Figure 1 and are described for all dates in Appendix A Table 7. This price series was inferred from billing data rather than directly observed, as discussed in Appendix A.

[11] In fact, we treat switching to an unpopular plan the same as quitting service, hence we also drop all remaining bills once a customer switches to an unpopular plan, even if they eventually switch back to a popular plan.



Plan	M_j	Q_j	p_j
Plan 0	\$14.99	0	\$0.11
Plan 1	\$34.99	380	\$0.45
Plan 2	\$44.99	653	\$0.40
Plan 3	\$54.99	875	\$0.40

Figure 1: Popular Plan Prices, Spring 2003.

3.2 Evidence for Stylized Facts

3.2.1 Three stylized facts relevant to modeling usage choices

Three features of the data are important to accurately model usage choices by customers of cellular phone service. First, consumers' usage choices are price sensitive. Second, consumers' usage choices are made while consumers are uncertain about the ex post marginal price. Third consumers are inattentive to the remaining balance of included minutes during the course of a billing cycle. These three stylized facts motivate our assumption that, rather than choosing a precise quantity, consumers choose calling thresholds and proceed to make all calls valued above the threshold.

Consumer price sensitivity is clearly illustrated by a sharp increase in calling volume on weekday evenings exactly when the off-peak period for free night and weekend calling begins (Figure 2). This is not simply a 9pm effect, as the increase occurs only on weekdays, and at 8pm for plans with early nights-and-weekends.[12]

Two pieces of evidence demonstrate consumer uncertainty about ex post marginal price. First, given clear sensitivity to marginal price, if consumers could anticipate whether they would be under their allowance (zero marginal price ex post) or over their allowance (35 to 45 cents per minute marginal price ex post) we would expect to see substantial bunching of consumers consuming their entire allowance but no more or less. Figure 3 shows there is no bunching, which is consistent with similar findings in the contexts of electricity consumption (Borenstein 2009) and labor supply

[12]For plans with free weeknight calling starting at 8pm, there is still a secondary increase in usage at 9pm (Figure 2 panel C). Restricting attention to outgoing calls made to land-lines (recipients for whom the cost of receiving calls was zero) almost eliminates this secondary peak (Figure 2 panel D). This suggests that the secondary peak is primarily due to calls to and from cellular numbers with 9pm nights (the most common time for free evening calling to begin) rather than a 9pm effect.

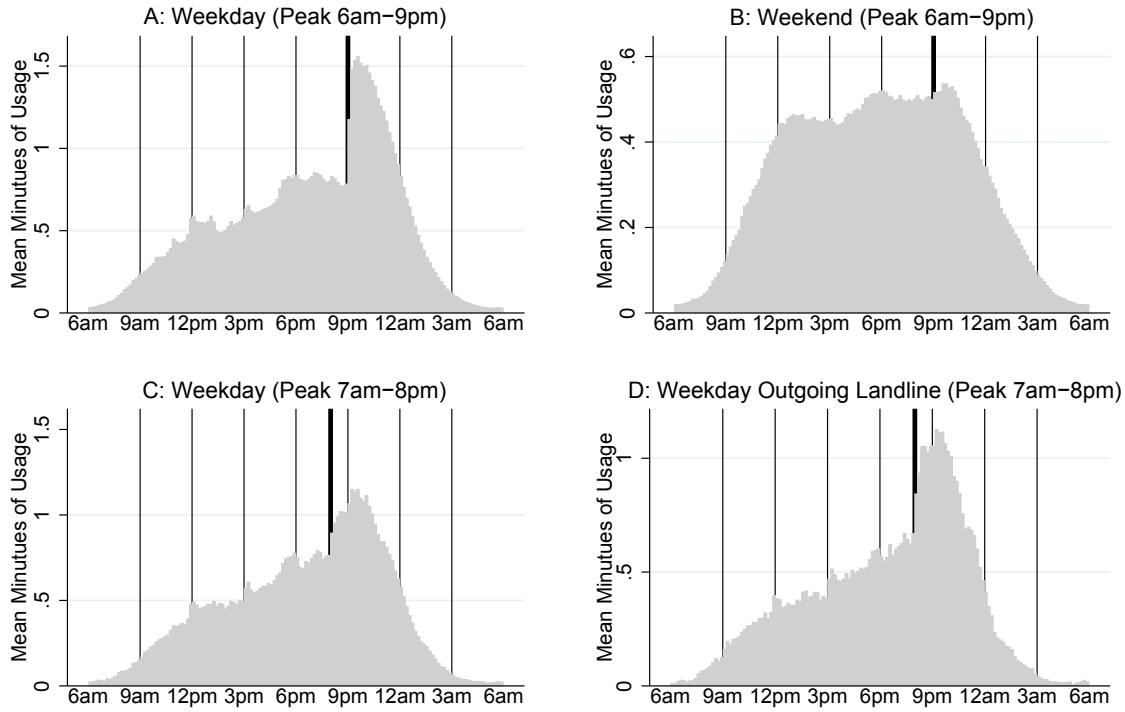


Figure 2: Daily usage patterns for subscribers with free nights and weekends. Top row: weekday (Panel A) and weekend (Panel B) usage patterns for subscribers with 6am-9pm peak hours. Bottom row: weekday usage patterns for subscribers with 7am-8pm peak hours. Panel C shows all weekday calling, while Panel D is restricted to outgoing calls to land-lines.

(Saez 2010). Second, consumers who anticipate being strictly under their allowance (zero marginal price ex post) should exhibit no price response at the commencement of off-peak hours. However, the sharp increase in calling at 9pm shown in Figure 2 persists even in months for which the peak allowance is under-utilized. These are natural consequences of usage choices made under uncertainty about ex post marginal price.

Now we turn to evidence that consumers are inattentive. If consumers are attentive to the remaining balance of included minutes during the billing cycle they should use this information to continually update their beliefs about the likelihood of an overage and a high marginal price ex post. Following an optimal dynamic program, an attentive consumer should (all else equal) reduce her usage later in the month following unexpectedly high usage earlier in the month. This should be true for any consumers who are initially uncertain whether they will have an overage in the current month. For these consumers, the high usage shock early in the month increases the likelihood of an overage, thereby increasing their expected ex post marginal price, and causing them to be more selective about calls. If calling opportunities arrived independently throughout the month, this

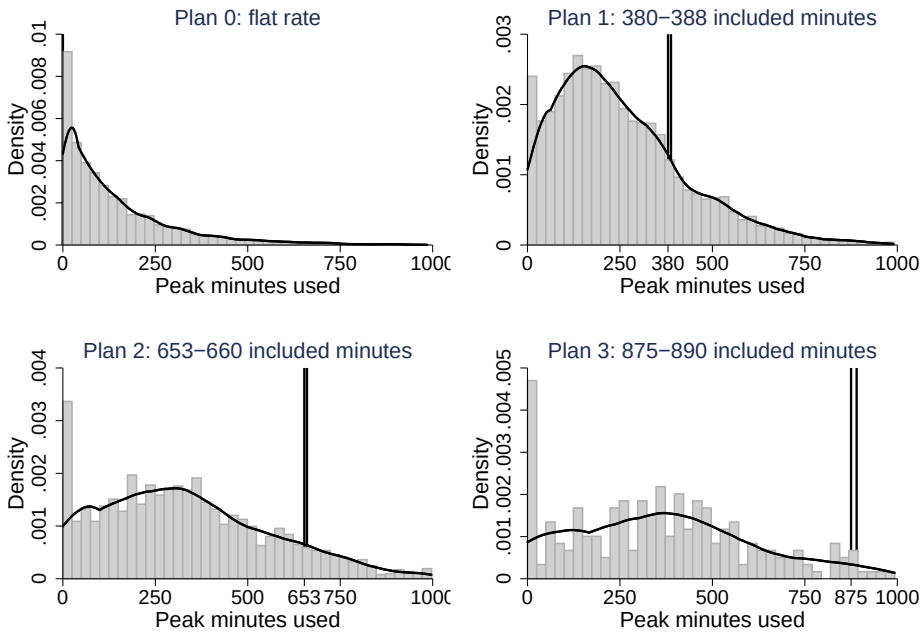


Figure 3: Usage densities for popular plans are constructed with 9,080, 5,026, 2,351, and 259 bills for plans 0-3 respectively. The sample for plans 1-3 is selected to only include bills for which in-network calls were costly and for which included peak minutes were within a narrow range, as indicated above each plot. Vertical lines bound the range of included free minutes for each plan.

strategic behavior by the consumer would lead to negative correlation between early and late usage within a billing period. However, looking for negative correlation in usage within the billing period is a poor test for this dynamic behavior because it is likely to be overwhelmed by positive serial correlation in taste shocks.

To test for dynamic behavior by consumers within the billing period, we use our data set of individual calls to construct both fortnightly and weekly measures of peak usage.[13] A simple regression of usage on individual fixed effects and lagged usage shows strong positive serial correlation. However, we take advantage of the following difference: Positive serial correlation between taste shocks in periods t and $(t-1)$ should be independent of whether periods t and $(t-1)$ are in the same or adjacent billing cycles. However, following unexpectedly high usage in period $(t-1)$, consumers should cut back usage more in period t if the two periods are in the same billing cycle. Thus by including an interaction effect between lagged usage and an indicator for the lag being in the same billing cycle as the current period, we can separate strategic behavior within the month from serial

[13]We divide each month into four weeks or two fortnights, and drop the extra 2-3 days between weeks 2 and 3.

correlation in taste shocks.

Table 1 shows a regression of log usage on lagged usage and the interaction between lagged usage and an indicator equal to 1 if period $(t-1)$ is in the same billing cycle as period t. We also include time and individual fixed effects and correct for bias induced by including both individual fixed effects and lags of the dependent variable in a wide but short panel (Roodman 2009). Reported analysis is for plan 1, the most popular three-part tariff. As expected, positive serial correlation in demand shocks leads to a positive and significant coefficient on lagged usage in the full sample (column 1) and most subsamples (columns 2-6). If consumers adjust their behavior dynamically within the billing cycle in response to usage shocks, then we expect the interaction effect to be negative. In the full sample (column 1) the interaction effect has a positive point estimate, but is not significantly different from zero. This suggests that consumers are not attentive to past usage during the course of the month.

Table 1: Dynamic usage pattern at fortnightly level.

	(1)	(2)	(3)	(4)	(5)	(6)
Overage Percentage	0-100%	0	1-29%	30-70%	71-99%	100%
$\ln(q_{t-1})$	0.649***	0.607***	0.535***	0.499***	-1.046	0.958***
	(0.0258)	(0.0529)	(0.0431)	(0.0683)	(1.065)	(0.0441)
SameBill*$\ln(q_{t-1})$	0.0133	0.0245	0.0193	-0.0149	-0.0837	3.685
	(0.0107)	(0.0183)	(0.0181)	(0.0222)	(1.180)	(4.745)
Observations	9068	3727	3218	1830	217	76
Number of id	386	167	130	87	11	6

Dependent variable $\ln(q_t)$. Standard errors in parentheses. Time and individual fixed effects.
Key: *** p<0.01, ** p<0.05, * p<0.1

Consumers who either never have an overage (43% of plan 1 subscribers) or always have an overage (3% of plan 1 subscribers) should be relatively certain what their ex post marginal price will be, and need not adjust calling behavior during the month. For instance, consumers who always make overages may only make calls worth more than the overage rate throughout the month. For such consumers we would expect to find no interaction effect, and this may drive the result when all consumers are pooled together as in our first specification. As a result, we divide consumers into groups by the fraction of times within their tenure that they have overages. We repeat our first specification for different *overage-risk groups* in Columns 2-6 of Table 1. The interaction effect is indistinguishable from zero in all overage risk groups. Moreover, in unreported analysis, more

flexible specifications that include nonlinear terms[14] and a similar analysis at the weekly rather than fortnightly level all estimate an interaction effect indistinguishable from zero. There is simply no evidence that we can find that consumers strategically cut back usage at the end of the month following unexpectedly high initial usage. We conclude that consumers are inattentive to their remaining balance of included minutes during the billing cycle.[15]

Our evidence for inattention is supported by Leider and Şahin's (2011) experimental work, which suggests that consumers who receive feedback about past usage do not follow an optimal dynamic program but instead use a constant calling threshold until all included minutes are used up and then adjust to the overage rate. This finding is consistent with our model of consumer behavior under our bill-shock counterfactual in which consumers are alerted when exceeding their allowance. In contrast, Yao et al. (2011) reject our static calling threshold model in favor of attentive dynamic behavior using Chinese cellular phone data.[16] The discrepancy between Yao et al.'s (2011) finding and our own may be due in part to the fact that, unlike consumers in our data, the Chinese consumers could check their minute balance. Moreover, results in all three papers can be reconciled by the fact that the financial incentives to pay attention were likely stronger for Chinese consumers than for American consumers and lab subjects.

3.2.2 Two stylized facts relevant to modeling plan choices

Two important features of the data are important to accurately model plan choice by cellular customers. First, while 29%-45% of contract choices are suboptimal ex post, consumers learn about their own usage levels over time and switch plans in response. Second, consumers' prior beliefs are biased: in the short run, before learning and switching plans, consumer plan-choice mistakes are predictable and can be exploited for profit. (We assume that consumers always make optimal plan-choices conditional on beliefs. When initial choices are suboptimal in a predictable way, we refer to consumers' prior beliefs as biased.)

Consumers switch plans. This may be in response to changes in tastes or prices but the pattern of switches shows that they are also made in response to learning. There are 1366 customers in our

[14] Average q_t will vary with expected marginal price, which is proportional to the probability of an overage. The probability of an overage in a billing period which includes periods t and $(t-1)$ increases nonlinearly in q_{t-1}. In one specification, we first fit a probit on the likelihood of an overage as a function of the first fortnights usage, and then used the estimated coefficients to generate overage probability estimates for all fortnights. We then included these (lagged) values as explanatory variables. In an alternative specification we added polynomial terms of lagged q_{t-1}.

[15] The finding is perhaps not surprising because service was resold by a university and, as a result, consumers could not contact the carrier to check minute balances.

[16] Yao et al. (2011) show that a scatter plot of cumulative weekly usage within a billing cycle against its lag is concave. In contrast, the relationship is linear in our data, which is consistent with our constant calling threshold.

data set, who we observe for an average of 12 months before either the data set ends or the customer quits.[17] Among all customers, 207 (15%) switch plans at least once, and 28 (2%) switch plans more than once, leading to a total of 246 plan switches. Of these switches, 85 (35%) are to plans that have either dropped in price or been newly introduced since the customer chose their existing plan. These switches could be motivated by price decreases rather than learning. However, the remaining 161 (65%) switches are to plans that are weakly more expensive than when the customer chose his or her existing plan. These switches must be due to learning or taste changes.

Not only do consumers switch plans, but they switch in the "right" direction. To substantiate this claim we make two calculations. First we calculate how much the customer would have saved had they signed up for the new plan initially, holding their usage from the original plan fixed. By this calculation, 60 to 61 percent of switches which can not be explained by price decreases saved customers money. (Switches that can not be explained by price decreases are those to plans which are weakly more expensive at the switching date than at the initial choice date.) Average savings, across money saving and money losing switches, are \$11.03 to \$15.44 per month.[18]

The savings estimates of \$11.03 to \$15.44 per month are underestimates because they do not take into account the fact that consumers can re-optimize usage choices upon switching plans. For instance, when switching to a plan with more included minutes consumers may optimally choose to talk more in response to the lower marginal price. An upper bound on the value of these additional calls is their price under the old plan. Hence our second calculation is the money that would have been lost had the customer not switched plans and remained on their original plan, again holding usage fixed. By this calculation average savings for switching are \$24.42 to \$31.84 per month, and 68 to 75 percent of switches saved money.[19] Hence consumers' expected benefit is between \$11.03 and \$31.84 per month when switching to plans that have not decreased in price since their previous choice, and 60 to 75 percent of switches are in the "right" direction.

---

[17] In our sample, 31 percent of customers are observed for more than 12 months. Standard cellular phone contracts often include switching costs (such as extension of commitment and delay of new phone subsidy) for switching plans prior to the expiry of one or two year contracts. In such a setting, more than 12 months of data would be needed to observe switching and learning. The students in our sample, however, could switch plans at any time and cancel after only three months, without any cost except hassle costs. As a result, we are able to observe active switching and learning over shorter time periods.

[18] We calculate bounds because we cannot always distinguish in-network and out-of-network calls. Both figures are statistically greater than zero at the 99% level. The 60-61 percent rates of switching in the "right" direction are statistically greater than 50 percent at the 95% level. This calculation is based on 98 of the 161 switches which can not be explained by price decreases. The remaining 63 switches occur so soon after the customer joins that there is no usage data prior to the switch that is not from a pro-rated bill.

[19] This calculation is based on 157 of the 161 switches which can not be explained by price decreases. The calculation cannot be made for the remaining 4 switches since there is no usage data following the switch that is not from a pro-rated bill. Figures are significant at the 99% confidence level.

In unreported analysis, additional evidence of learning is that: (1) the likelihood of switching declines with tenure, and (2) the likelihood of switching to a larger plan increases after an overage. Narayanan et al. (2007) estimate that consumers in the Kentucky experiment learn to switch up from overuse faster than they learn to switch down from underuse. In the context of retail banking, Ater and Landsman's (2011) results suggest that the asymmetry could be large enough that banking customers' tendency too choose overly large plans grows overtime through switching. For simplicity, we implement symmetric learning in our structural model.

Table 2: Predictable Customer Mistakes Yield Savings Opportunities

	First Opportunity	Second Opportunity
Dates	10/02-8/03	9/03 onwards
Enrollment Change	plan 1-3 $\rightarrow$ plan 0	plan 1 $\rightarrow$ plan 2
Affected Customers	251 (34%)	445 (55%)
Savings		
Total	\$20,840 (47%)	\$7,942 (28%)
Per Affected Bill	\$8.76	\$2.64
Per Affected Cust.	\$83.03 (149%)	\$17.85 (46%)

The University acts as a reseller and could bill students for their chosen plan, sign them up for an alternative plan, and save the difference in charges. These plan-level savings opportunities indicate that consumers choose overly risky plans (overconfidence or projection bias). Savings estimates are a lower bound because we cannot always distinguish in and out-of-network calls.

The presence of ex post mistakes alone shows only that consumers face uncertainty ex ante at the time of plan choice. However, ex post mistakes are not only present, they are also predictable given an individual's initial plan choice and population usage data. This implies that consumers' prior beliefs are biased and differ from average posteriors. Two plan-level savings opportunities demonstrate that customer mistakes are predictable and show how such predictability can be exploited by firms. (The first savings opportunity is an extension of that documented in Grubb (2009).) The university acts as a reseller and charges students a fixed five dollar fee per month to cover administrative costs. Although the university did not do so, they could have billed students based on the terms of their chosen calling plan, but signed them up for a predictably cheaper plan and saved the difference in charges. Table 2 illustrates two substantial opportunities. In the 2002-2003 academic year, when plan 0 offered free off-peak calling, by signing the 248 students who selected plans 1-3 up for plan 0, the university would have saved at least \$20,731, or \$83.59 per affected student. In the following year, the cellular company closed this opportunity by ending free off-peak calling on plan 0. However, an alternative was to sign up the 439 students who chose

plan 1 onto plan 2, which would have saved at least \$7,934, or \$18.07 per affected student. These plan-level savings opportunities indicate that consumers choose overly risky plans (overconfidence or volatility bias).[20]

4 Model

At each date t, consumer i first chooses a plan j and then chooses peak and off-peak quantities summarized by the vector $\mathbf{q}_{it} = (q_{it}^{pk}, q_{it}^{op})$. (The text suppresses the distinction between in-network and out-of-network calling, which is covered in Appendix C.) Total billable minutes for plan j are

$$q_{itj}^{billable} = q_{it}^{pk} + OP_j q_{it}^{op},$$

where OP_j is an indicator variable for whether plan j charges for off-peak usage. At the end of period t, consumer i is charged

$$P_j\left(\mathbf{q}_{it}\right) = M_j + p_j \max\{0, q_{itj}^{billable} - Q_j\},$$

where pricing plan j has monthly fee M_j, included allowance Q_j, and overage rate p_j.

We assume consumers are risk neutral, consumers have quasi-linear utility, and peak and off-peak calls are neither substitutes nor complements.[21] Consumer i's money-metric utility in month t from choosing plan j and consuming $\mathbf{q}_{it}$ units is

$$u_{itj} = \sum_{k \in \{pk, op\}} V\left(q_{it}^k, \theta_{it}^k\right) - P_j\left(\mathbf{q}_{it}\right) + \frac{1}{\alpha}\eta_{itj},$$

where

$$V\left(q_{it}^k, \theta_{it}^k\right) = \frac{1}{\beta}\left(\theta_{it}^k \ln\left(q_{it}^k / \theta_{it}^k\right) - q_{it}^k\right)$$

is the value from category $k \in \{pk, op\}$ calling, which depends on a pair of non-negative taste-shocks $\boldsymbol{\theta}_{it} = (\theta_{it}^{pk}, \theta_{it}^{op})$, and η_{itj} is an i.i.d. logit error.[22] The marginal value of a dollar is normalized to

[20] Aggregate and conditional mean biases could explain one or other plan-level savings opportunity but only overconfidence and volatility bias can simultaneously explain both savings opportunities. Note that the first savings opportunity is robust to dropping the top 30 percent of customers with the highest average savings, while the second savings opportunity is robust to dropping the top 2 percent of customers.

[21] In reality, consumers likely do delay calls until off-peak periods. Our assumption ruling out such substitution should not bias our final results. In particular, as off-peak calling is typically free and is exogenously so in our counterfactual simulations, whether peak calls are foregone entirely or shifted off-peak does not effect firm revenues or peak-pricing. Moreover, in either case, foregone peak calls carry a social cost captured in our welfare estimates.

[22] We model consumers' choice between the four most popular pricing plans (plans 0-3), comparable plans from other

one, $1/\alpha$ scales the logit error variance, and β is a price coefficient that determines how sensitive calling choices are to the marginal price of an additional minute of calling time. Our choice of functional form for $V\left(q_{it}^k, \theta_{it}^k\right)$ implies that the taste shock θ_{it}^k enters demand multiplicatively, as discussed below.

4.1 Quantity Choices

Recognizing that consumers are uncertain about the ex post marginal price when making usage choices from three-part tariffs is a key feature of our model and where we take a new approach (also suggested independently by Borenstein (2009)). We assume that at the start of billing period t, consumer i is uncertain about her period t taste shock $\boldsymbol{\theta}_{it}$. She first chooses a plan j and then chooses a calling threshold vector $\mathbf{v}_{itj}^* = (v_{itj}^{pk}, v_{itj}^{op})$ based on chosen plan terms and her beliefs about the distribution of $\boldsymbol{\theta}_{it}$. During the course of the month, the consumer is inattentive and does not track usage but simply makes all category-k calls valued above v_{itj}^k. Over the course of the month, for $k \in \{pk, op\}$ this cumulates to the choice:

$$q_{it}^k = q(v_{itj}^k, \theta_{it}^k) = \theta_{it}^k \hat{q}(v_{itj}^k), \tag{1}$$

where $\hat{q}\,(v) = 1/\,(1+\beta v)$ and $\hat{q}\,(0) = 1$.[23]

The interpretation is that θ_{it}^k is the volume of category-k calling opportunities that arise and $\hat{q}(v)$ is the fraction of those calling opportunities worth more than v per minute. Timing is summarized in Figure 4. Figure 5 shows the calling threshold v_{itj}^{pk} and resulting consumption choice $\theta_{it}^{pk}\hat{q}(v_{itj}^{pk})$ in relation to a consumer's realized inverse demand curve for calling minutes, $V_q(q_{it}^{pk}, \theta_{it}^{pk})$.

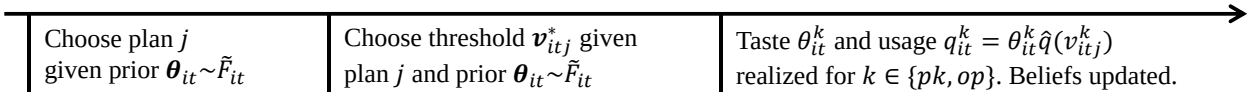


Figure 4: Model Time Line

carriers, and an outside option. For plans other than the four popular university plans, the logit error η_{itj} has a clear economic interpretation: it includes all unmodeled plan heterogeneity including network quality, available phones, and roaming charges. Within the four popular plans, the logit error η_{itj} has no satisfactory economic interpretation, as these plans only differ in price, and in the complete model we capture all the dimensions on which prices differ. All initial plan choices could be explained without including the logit error, but they are required to explain switches that appear to be in the "wrong" direction.

[23]The fact that demand is multiplicative in θ_{it}^k follows from the assumption that $V\left(q_{it}, \theta_{it}^k\right)$ can be expressed as $V\left(q_{it}, \theta_{it}^k\right) = \theta_{it}^k \hat{V}\left(q_{it}/\theta_{it}^k\right)$ for some function $\hat{V}$. In this case, $\hat{V}\,(x) = (\ln x - x)/\beta$. The fact that $\hat{q}(0) = 1$ simply reflects the chosen normalization of θ_{it}^k.

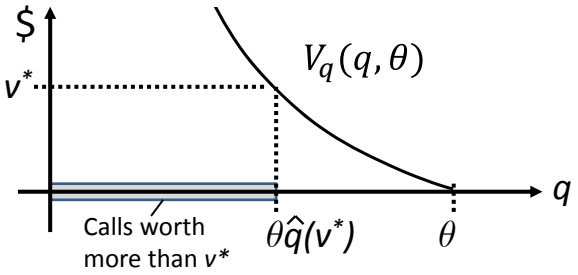


Figure 5: Inverse Demand Curve and Calling Threshold

Making all peak calls valued above the constant threshold v^*_{itj} is the optimal strategy of an inattentive consumer who does not track usage within the current billing cycle and hence cannot update his beliefs about the likelihood of an overage within the current billing cycle. (It is analogous to an electricity consumer setting a thermostat rather than choosing a quantity of kilowatt hours.)

When marginal price is constant, a consumer's optimal calling threshold is simply equal to the marginal price. Thus for plan zero, which charges 11 cents per minute for all billable calls, $\mathbf{v}^*_{itj} = (0.11, 0.11OP_j)$. Further, $v^{op}_{itj} = 0$ for plans 1-3 because they offer free off-peak calling.

Conditional choosing one of plans 1-3, which include free off-peak calling and an allowance of peak minutes, consumer i chooses her period t peak-calling threshold v^{pk}_{itj} to maximize her expected utility conditional on her period t information $\Im_{it}$. Given allowance Q_j, overage rate p_j, and multiplicative demand (equation (1)), the optimal threshold (derived in Appendix B.1) is uniquely characterized by equation (2):

$$v^{pk}_{itj} = p_j \Pr\left(\theta^{pk}_{it} \geq Q_j/\hat{q}(v^{pk}_{itj}) \mid \Im_{it}\right) \frac{E\left[\theta^{pk}_{it} \mid \theta^{pk}_{it} \geq Q_j/\hat{q}(v^{pk}_{itj});\ \Im_{it}\right]}{E\left[\theta^{pk}_{it} \mid \Im_{it}\right]}. \tag{2}$$

The threshold v^{pk}_{itj} will be between zero and the overage rate p_j.[24]

Note that choosing threshold v^{pk}_{itj} is equivalent to choosing a target peak-calling quantity $q^T_{it} \equiv E[\theta^{pk}_{it}]\hat{q}(v^{pk}_{itj})$, which is implemented with error $(\theta^{pk}_{it} - E[\theta^{pk}_{it}])\hat{q}(v^{pk}_{itj})$. Importantly, consumers are aware of their inability to hit the target precisely and take this into account when making their threshold/target choice.

[24]Equation (2) may seem counter-intuitive, because the optimal v^{pk}_{itj} is greater than the expected marginal price, $p_j \Pr(q(v^{pk}_{itj}, \theta^{pk}_{it}) > Q_j \mid \Im_{it})$. This is because the reduction in consumption from raising v^{pk}_{itj} is proportional to θ^{pk}_{it}. Raising v^{pk}_{itj} cuts back on calls valued at v^{pk}_{itj} more heavily in high demand states when they cost p_j and less heavily in low demand states when they cost 0.

4.2 Plan Choices

We model consumers' choice between the four most popular pricing plans (plans 0-3), comparable AT&T, Cingular, and Verizon plans (Sprint offered no local plans), and an outside option which incorporates all other plans. We adopt Ching, Erdem and Keane's (2009) consideration set model by assuming that consumers make an active choice with exogenous probability P_C and keep their current plan with probability $(1 - P_C)$. We use the frequency of failures to switch away from dominated plans to identify P_C.[25]

Customer i's perceived expected utility from choosing plan j at date t is

$$U_{itj} = E\left[\sum_{k\in\{pk,op\}} V\left(q(v^k_{itj},\theta^k_{it}),\theta^k_{it}\right) - P_j\left(\mathbf{q}(\mathbf{v}^*_{itj},\boldsymbol{\theta}_{it})\right) \mid \Im_{it}\right] + \frac{1}{\alpha}\eta_{itj}, \tag{3}$$

and from choosing the outside option is $U_{it0} = O + \eta_{it0}$. The parameter O will be identified from the frequency at which consumers leave the data set. Conditional on making an active choice, a consumer's consideration set includes plans offered by her current provider, the outside option, and plans from a randomly selected alternative carrier.[26] Consumers myopically[27] choose the plan (or outside option) from their consideration set that maximizes expected utility in the current period.

4.3 Distribution of Tastes

We assume that the non-negative taste-shocks which determines usage are latent taste shocks censored at zero:

$$\theta^k_{it} = \begin{cases} 0 & \tilde{\theta}^k_{it} < 0 \\ \tilde{\theta}^k_{it} & \tilde{\theta}^k_{it} \geq 0 \end{cases}, \; k \in \{pk, op\}.$$

---

[25] When prices fall consumers often do not switch away from their existing plans even when they are now dominated by plans on the current menu. For instance, most consumers paying \$54.99 for 890 minutes on plan 3 do not switch to plan 2 during the one month promotion in April 2004 when it offered 1060 minutes for only \$44.99. We believe this is because consumers who are not actively making a plan choice do not find out about the price cuts.

[26] We avoid including all plans in the consideration set to reduce computational time.

[27] We assume learning is independent of plan choice, so there is no value to experimentation with an alternative plan. Nevertheless, myopic plan choice is not optimal for several reasons. First, when a consumer is currently subscribed to a plan that is no longer offered (and is not dominated) there is option value to not switching, since switching plans will eliminate that plan from future choice sets. Second, if $P_C < 1$, a forward looking consumer would tend to discount her current period logit-error η_{it}. Third, if $P_C < 1$, a forward looking consumer should anticipate that her current plan choice may persist in the future but her future calling threshold choices v^* will improve as she learns about her type μ_i. This consideration makes plans 1 and 2 marginally more attractive relative to plans 0 and 3 but the effect is not large. We ignore these issues for tractability.

We assume that the latent shock $\tilde{\theta}_{it}^{k}$ is normally distributed and that consumers observe its value even when censored. This adds additional unobserved heterogeneity to the model but preserves tractable Bayesian updating. Censoring makes zero usage a positive likelihood event, which is important since it occurs for 10% of plan 0 observations.

Usage choices in the data are strongly serially-correlated conditional on customer-plan and date fixed effects. We therefore incorporate simple serial-correlation into our model by assuming that the latent shock $\tilde{\boldsymbol{\theta}}_{it}$ follows a stationary AR1 process with a bivariate normal innovation,

$$\tilde{\boldsymbol{\theta}}_{it} = \boldsymbol{\mu}_i + \varphi \tilde{\boldsymbol{\theta}}_{i,t-1} + \boldsymbol{\varepsilon}_{it},$$

where $\boldsymbol{\mu}_i$ is customer i's true type, φ is the common serial coefficient, and $\boldsymbol{\varepsilon}_{it} \sim N(0, \boldsymbol{\Sigma}_\varepsilon)$ is the normally-distributed mean-zero innovation with variance-covariance matrix

$$\boldsymbol{\Sigma}_\varepsilon = \begin{bmatrix} (\sigma_\varepsilon^{pk})^2 & \rho_\varepsilon \sigma_\varepsilon^{pk} \sigma_\varepsilon^{op} \\ \rho_\varepsilon \sigma_\varepsilon^{pk} \sigma_\varepsilon^{op} & (\sigma_\varepsilon^{op})^2 \end{bmatrix}.$$

(We assume AR(1) rather than AR(k) for simplicity.) Consumers' true types, $\boldsymbol{\mu}_i = (\mu_i^{pk}, \mu_i^{op})$, are normally distributed across the population as described below.

4.4 Beliefs and Learning

Estimation of consumer beliefs and learning is focused on a single dimension of usage: total peak-calling. We make this restriction because plans 1-3 always offer free off-peak calling and hence the choice data are not rich enough to allow us to identify beliefs about off-peak calling. For simplicity, we assume that while consumers are learning about their peak type μ_i^{pk} over time, there is no learning about off-peak demand because consumers know their off-peak types μ_i^{op}.[28]

We assume the serial-correlation coefficient φ is known by all consumers. While taste innovations $\boldsymbol{\varepsilon}_{it}$ have variance-covariance $\boldsymbol{\Sigma}_\varepsilon$, consumers believe the variance-covariance matrix is

$$\tilde{\boldsymbol{\Sigma}}_\varepsilon = \begin{bmatrix} (\tilde{\sigma}_\varepsilon^{pk})^2 & \rho_\varepsilon \tilde{\sigma}_\varepsilon^{pk} \sigma_\varepsilon^{op} \\ \rho_\varepsilon \tilde{\sigma}_\varepsilon^{pk} \sigma_\varepsilon^{op} & (\sigma_\varepsilon^{op})^2 \end{bmatrix},$$

where $\tilde{\sigma}_\varepsilon^{pk} = \delta_\varepsilon \sigma_\varepsilon^{pk}$ and $\delta_\varepsilon > 0$. If $\delta_\varepsilon = 1$, then consumers' perceptions match reality. If $\delta_\varepsilon < 1$, then consumers underestimate the volatility of their peak tastes from month-to-month and exhibit

[28] This assumption does not effect our endogenous-price counterfactual simulations because we assume free off-peak calling.

volatility bias. If $\delta_\varepsilon < 1$, then consumers will predictably choose too risky plans and overreact to past usage when deciding whether or not to switch plans.[29] Consumer beliefs about the variance of off-peak tastes and the correlation between peak and off-peak tastes are both correct.

Consumers learn about their own peak-type μ_i^{pk} over time. At date t, consumer i believes that μ_i^{pk} is normally distributed with mean $\tilde{\mu}_{i,t}^{pk}$ and variance $\tilde{\sigma}_t^2$: $\mu_i^{pk}|\Im_{i,t} \sim N(\tilde{\mu}_{i,t}^{pk}, \tilde{\sigma}_t^2)$. At the end of each billing period, usage q_{it}^{pk} is realized and consumers can infer $\theta_{it}^{pk} = q_{it}^{pk}/\hat{q}(v_{itj}^{pk})$. When $q_{it}^{pk} = \theta_{it}^{pk} = 0$, we assume that consumers can observe the latent taste shock $\tilde{\theta}_{it}^{pk}$. The latent shock provides an unbiased normal signal about μ_i^{pk} and consumers update beliefs according to Bayes rule (see Appendix B.3).[30] Over time consumers learn their own types: $\tilde{\mu}_{i,t}^{pk}$ converges to μ_i^{pk} and $\tilde{\sigma}_t^2$ converges to zero.

Consumers' plan choices and threshold choices depend on beliefs about the distribution of tastes $\boldsymbol{\theta}_{it}$. When choosing a plan and a usage threshold for the first time, consumers believe:

$$\tilde{\theta}_{i1}^{pk} \sim N\left(\frac{\tilde{\mu}_{i1}^{pk}}{1-\varphi}, \tilde{\sigma}_{\theta 1}^2\right), \tag{4}$$

where

$$\tilde{\sigma}_{\theta 1}^2 = \frac{\tilde{\sigma}_1^2}{(1-\varphi)^2} + \frac{(\tilde{\sigma}_\varepsilon^{pk})^2}{1-\varphi^2}. \tag{5}$$

In all later periods $t > 1$, when consumers can condition on $\tilde{\theta}_{i,t-1}^{pk}$, beliefs are:

$$\tilde{\theta}_{it}^{pk} \mid \Im_{it} \sim N\left(\tilde{\mu}_{it}^{pk} + \varphi\tilde{\theta}_{it-1}^{pk}, \tilde{\sigma}_t^2 + (\delta_\varepsilon\sigma_\varepsilon^{pk})^2\right).$$

Following a month with surprisingly high usage, consumer i's belief about the distribution of demand in the following month increases for two reasons. First the consumer increases his estimate of his type ($\tilde{\mu}_{i,t+1}^{pk} > \tilde{\mu}_{it}^{pk}$), and second he knows that his demand is positively correlated over time. In the standard model the only behavior change that might result is a switch to a larger plan. In our model, a consumer might also switch to a larger plan but, conditional on not switching, would cut back on usage by choosing a higher calling threshold ($v_{i,t+1}^{pk} > v_{i,t}^{pk}$) and being more selective about calls.

---

[29] For tractability, we assume that consumers learn about means but not variances, so volatility bias is persistent.

[30] In fact, given our assumption that consumers know μ_i^{op}, consumers can also infer ε_{it}^{op} from off peak usage which is informative about μ_i^{pk} because it is correlated with ε_{it}^{pk}. We assume consumers only update beliefs using θ_{it}^{pk} and not ε_{it}^{op}. This choice is conservative in the sense that our finding that consumers respond to data too little is biased downwards. It is also realistic for two reasons. First, consumers are unlikely to pay attention to off-peak usage when they are on contract with free off-peak calls. Second, we only assume consumers know μ_i^{op} for simplicity as we cannot identify off-peak beliefs. In reality, consumers are unlikely to know μ_i^{op} so cannot actually infer ε_{it}^{op}.

4.5 Priors

Each customer is characterized by the individual specific triple $\{\tilde{\mu}_{i1}^{pk}, \mu_i^{pk}, \mu_i^{op}\}$. Together with the population parameter $\tilde{\sigma}_1^2$, this triple specifies each customer's true type $\boldsymbol{\mu}_i$ and prior belief $\mu_i^{pk} \sim N(\tilde{\mu}_{i1}^{pk}, \tilde{\sigma}_1^2)$. (Consumers are assumed to know their own off-peak types.) The population is described by the joint distribution of $\{\tilde{\mu}_{i1}^{pk}, \mu_i^{pk}, \mu_i^{op}\}$, which we assume is a trivariate normal distribution. We outline the nine parameters of the distribution at the end of the section. By not restricting any of these parameters, our formulation allows for consumer beliefs about their peak types to be biased in three ways, as described below.

First, aggregate mean bias can arise when the average point estimate is too low or too high. We define μ_0^{pk} and $\tilde{\mu}_0^{pk}$ to be the population averages of true peak types μ_i^{pk} and prior point estimates $\tilde{\mu}_{i1}^{pk}$, respectively. A typical assumption (perhaps labeled rational expectations) is that $\mu_0^{pk} = \tilde{\mu}_0^{pk}$, which implies that an average individual's initial point estimate is an unbiased estimate of her true type. We do not impose this assumption. If $b_1 \equiv \tilde{\mu}_0^{pk} - \mu_0^{pk} \neq 0$, then there is aggregate mean bias and consumers will predictably choose plans which are too small ($b_1 < 0$) or too large ($b_1 > 0$).

Second, overconfidence can arise when the precision of consumers' beliefs about their type is miscalibrated. In our notation, $\sigma^2_{\mu^{pk}}$ is the conditional variance of true peak types in the population, $Var(\mu_i^{pk} \mid \tilde{\mu}_{i1}^{pk})$, and $\tilde{\sigma}_1$ is consumers' uncertainty about their peak type. A typical assumption (perhaps labeled rational expectations) is that $\tilde{\sigma}_1 = \sigma_{\mu^{pk}}$.[31] We do not impose this assumption either. If $\delta_\mu \equiv \tilde{\sigma}_1/\sigma_{\mu^{pk}} < 1$ then consumers exhibit overconfidence: they underestimate their own uncertainty about their type μ_i^{pk}. Overconfident consumers, like those with volatility bias, will predictably choose overly risky plans. However, in contrast to those with volatility bias, they will under-react to past usage when making plan switching decisions. Grubb's (2009) analysis is static, so could not distinguish between overconfidence and volatility bias, but found that customers do choose overly risky plans, so exhibit either overconfidence, volatility bias, or both.

Third, conditional mean bias can arise when consumers over or under react to their own private information, forming individual point estimates, $\tilde{\mu}_{i1}^{pk}$, that differ from the population average, $\tilde{\mu}_0^{pk}$, too much or too little. Conditional on $\tilde{\mu}_{i1}^{pk}$, we write the population average of μ_i^{pk} as $E\left[\mu_i^{pk} \mid \tilde{\mu}_{i1}^{pk}\right] = \mu_0^{pk} + \psi^{pk}(\tilde{\mu}_{i1}^{pk} - \tilde{\mu}_0^{pk})$.[32] Letting $b_2 \equiv 1 - \psi^{pk}$, we have:

$$\tilde{\mu}_{i1}^{pk} - E\left[\mu_i^{pk} \mid \tilde{\mu}_{i1}^{pk}\right] = b_1 + b_2(\tilde{\mu}_{i1}^{pk} - \tilde{\mu}_0^{pk}). \tag{6}$$

[31] An alternative rational-expectations benchmark discussed in Appendix B.4 would be $\tilde{\sigma}_1^2 = Var(\mu_i^{pk} \mid \tilde{\mu}_{i1}^{pk}, \mu_i^{op})$. Using this benchmark would affect the description of bias but not its economic consequences: it would not alter our evaluation of bill-shock regulation or any welfare results outside of our de-biasing counterfactual simulations.

[32] Implicitly this defines ψ^{pk} as $Cov(\mu_i^{pk}, \tilde{\mu}_{i1}^{pk})/Var(\tilde{\mu}_{i1}^{pk})$.

A typical assumption is that $\tilde{\mu}_{i1}^{pk} = E[\mu_i^{pk} \mid \tilde{\mu}_{i1}^{pk}]$, or $b_1 = b_2 = 0$, which implies that all individuals' initial point estimates are unbiased estimates of their true types.[33] (Assuming $b_1 = 0$ only makes the weaker restriction that the average individual's point estimate is unbiased.) The parameter b_2 measures the amount of conditional mean bias in the population. If $b_2 > 0$ then point estimates differ too much from the population average and consumers predictably choose plans which are too extreme. If $b_2 < 0$ then point estimates differ too little from the population average and consumers predictably choose plans which are too moderate.

For completeness, we finish by describing the joint normal distribution of $\{\tilde{\mu}_{i1}^{pk}, \mu_i^{pk}, \mu_i^{op}\}$ in two parts. First, the marginal distribution of initial point estimates $\tilde{\mu}_{i1}^{pk}$ is normal with mean $\tilde{\mu}_0^{pk}$ and variance $\tilde{\sigma}^2_{\mu^{pk}}$. Second, conditional on the point estimate $\tilde{\mu}_{i1}^{pk}$, the population distribution of true types $\boldsymbol{\mu}_i$ is normal with mean $\boldsymbol{\mu}_0 + \boldsymbol{\psi}(\tilde{\mu}_{i1}^{pk} - \tilde{\mu}_0^{pk})$ and variance matrix $\boldsymbol{\Sigma}_\mu$. The vectors $\boldsymbol{\mu}_0$ and $\boldsymbol{\psi}$ are defined as (μ_0^{pk}, μ_0^{op}) and (ψ^{pk}, ψ^{op}), respectively; the diagonal elements of $\boldsymbol{\Sigma}_\mu$ are $\sigma^2_{\mu^{pk}}$ and $\sigma^2_{\mu^{op}}$, while the off-diagonal is $\rho_\mu \sigma_{\mu^{pk}} \sigma_{\mu^{op}}$, where ρ_μ is the conditional correlation of μ_i^{pk} and μ_i^{op}. Note that the joint distribution of true types and priors we describe can naturally be generated from the marginal distribution of true types, a common prior, and an unbiased signal that accounts for aggregate uncertainty. This is the presentation adopted by Goettler and Clay (2011).

5 Identification

Parameters can be categorized into four groups: (1) parameters governing beliefs ($\tilde{\mu}_0^{pk}$, $\tilde{\sigma}_{\mu^{pk}}$, $\tilde{\sigma}_1$, and $\tilde{\sigma}_\varepsilon^{pk}$), (2) the true (conditional) distribution of tastes ($\boldsymbol{\mu}_0$, $\boldsymbol{\psi}$, $\boldsymbol{\Sigma}_\mu$, $\boldsymbol{\Sigma}_\varepsilon$, and φ), (3) the price coefficient β, and (4) parameters related to switching and quitting (P_C, α, and O). Broadly speaking, plan choices identify beliefs, the distribution of actual usage identifies the distribution of true tastes, and changes in usage in response to the discontinuous change in marginal price between peak and off-peak hours identify the price coefficient β. Finally, the rate of switching away from dominated plans, the rate of switching in the "wrong" direction, and the rate of quitting identify, respectively, the active choice probability P_C, the logit error weight $1/\alpha$, and the outside option O.

5.1 Price Coefficient

If consumers' chosen thresholds ($\mathbf{v}_{it}^*$) were known, the price coefficient β could be inferred from marginal price variation and the induced variation in $\hat{q}(v_{it}^k)$.[34] Unfortunately, we require β to

[33] An alternative benchmark discussed in Appendix B.4 would be $\tilde{\mu}_{i1}^{pk} = E[\mu_i^{pk} \mid \tilde{\mu}_{i1}^{pk}, \mu_i^{op}]$. See foonote 31.

[34] For instance, there is one clean experiment in the data in which existing plan 1 subscribers were automatically upgraded from 280 free minutes to 380 free minutes and increased their usage in response by an average of 53 minutes.

calculate $\mathbf{v}_{it}^*$. We circumvent this problem by relying on a source of marginal price variation for which v_{it}^* is known. Prior to fall 2003, v_{it}^* is 11 cents during peak hours and 0 cents during off-peak hours for plan 0 subscribers.

Although prices in the model depend only on total peak and total off-peak calling, we additionally break out the share of calling demand for weekday outgoing-calls to landlines immediately before and after 9pm to help identify the price coefficient. The shock $\mathbf{r}_{it}^{9pm} = (r_{it}^{9pk}, r_{it}^{9op}) \in [0,1]^2$ captures the share of peak and off-peak calling demand that is within 60 minutes of 9pm on a weekday and is for an outgoing call to a landline. The distribution of r_{it}^k for $k \in \{9pk, 9op\}$ is a censored normal,

$$\begin{aligned} \tilde{r}_{it}^k &= \alpha_i^k + e_{it}^{r,k} \\ r_{it}^k &= \begin{cases} 0 & \text{if} \quad \tilde{r}_{it}^k \leq 0 \\ \tilde{r}_{it}^k & \text{if} \quad 0 < \tilde{r}_{it}^k < 1 \\ 1 & \text{if} \quad \tilde{r}_{it}^k \geq 1 \end{cases}, \end{aligned}$$

where α_i^k is unobserved heterogeneity and $e_{it}^{r,k}$ is a mean-zero shock normally distributed with variance $(\sigma_e^k)^2$ independent across i, t, and k. We assume that α_i^{9pk} is normally distributed in the population with mean μ_α^{9pk} and variance $(\sigma_\alpha^{9pk})^2$.

Our identifying assumption for the price coefficient is that consumer i's expected outgoing calling demand to landlines on weekdays is the same between 8:00pm and 9:00pm as it is between 9:00pm and 10:00pm:

$$E\left[r_{it}^{9pk}\right] E\left[\theta_{it}^{pk}\right] = E\left[r_{it}^{9op}\right] E\left[\theta_{it}^{op}\right]. \tag{7}$$

In other words, we assume that the increase in observed calling to landlines on weekdays immediately after off-peak begins at 9pm is a price effect rather than a discontinuous increase in demand at 9pm.[35] As a result, equation (7) implicitly defines α_i^{9op} as a function of α_i^{9pk} and other parameters.

Given plan 0 pricing prior to fall 2003, $\theta_{it}^{op} = q_{it}^{op}$ and $\theta_{it}^{pk} = q_{it}^{pk}\left(1 + 0.11\beta\right)$. Moreover, the pre and post 9pm calling shares are always observed because calling thresholds are constant within peak and within off-peak hours: $r_{it}^{9op} = q_{it}^{9op}/q_{it}^{op}$ and $r_{it}^{9pk} = q_{it}^{9pk}/q_{it}^{pk}$. Thus equation (7) can be

(The 95% confidence interval on this increase is 26-81 minutes.) However, without knowing how consumer thresholds were affected by the price change, this does not identify β.

[35] We focus on calls to landlines because the other party to the call pays nothing both before and after 9pm. The assumption would be unreasonable for calls to or from cellular numbers since such calling opportunities increase at 9pm when the calls become cheaper for the other party and the other party is more likely to call or answer.

solved for β as a function of moments of the data:

$$\beta = \frac{100}{11}\left(\frac{E\left[q_{it}^{9op}/q_{it}^{op}\right]E\left[q_{it}^{op}\right]}{E\left[q_{it}^{9pk}/q_{it}^{pk}\right]E\left[q_{it}^{pk}\right]} - 1\right).$$

5.2 Serial Correlation

Data prior to fall 2003 identifies the AR1 coefficient φ. During this period, all plans offered free nights-and-weekends so that we observe

$$q_{it}^{op} = \theta_{it}^{op} = \mu_i^{op} + \varphi\theta_{it-1}^{op} + \epsilon_{it}^{op}. \tag{8}$$

The argument follows the identification argument for the parameters of a linear regression model with person level fixed effects and a lagged dependent variable. By taking the first difference of equation (8), we remove the impact of the fixed effect μ_i^{op}. Then φ can be estimated using past values of θ_{it}^{op} as instruments, as in Blundell and Bond (1998).

5.3 Beliefs

Next, consider identification of consumers' prior beliefs from plan choices. Choice data are quite informative about beliefs about peak usage, as illustrated by Figure 6, but relatively uninformative about beliefs about off-peak usage. Hence we assume consumers know their own off-peak taste distribution (including μ_i^{op} and σ_ε^{op}). Prior to fall 2003, when off-peak calling is free, an individual consumer's initial plan choice depends only on β, η_{i1}/α, and her beliefs about θ_{i1}^{pk} described by $\tilde{\mu}_{i1}^{pk}/(1-\varphi)$ and $\tilde{\sigma}_{\theta 1}$. Thus initial plan-choice shares depend only on α, β, φ, $\tilde{\sigma}_{\theta 1}$, and the population distribution of $\tilde{\mu}_{i1}^{pk}$, described by $\tilde{\mu}_0^{pk}$ and $\tilde{\sigma}^2_{\mu^{pk}}$. Parameters φ and β are already identified. For transparency of the argument, we begin by considering a restricted model that excludes logit errors ($1/\alpha = 0$). Initial plan choice shares identify the remaining parameters $\tilde{\mu}_0^{pk}$, $\tilde{\sigma}^2_{\mu^{pk}}$, and $\tilde{\sigma}_{\theta 1}$. Finally, the learning rate separately identifies $\tilde{\sigma}_1$ and $\tilde{\sigma}_\varepsilon$ from $\tilde{\sigma}_{\theta 1}$. Initial choice shares in post fall 2003 data also aid identification, but require a more complicated argument involving beliefs about off-peak tastes.

Absent the logit-error, initial plan choices place bounds on each individual's prior beliefs about the mean ($\tilde{\mu}_{i1}^{pk}/(1-\varphi)$) and variance ($\tilde{\sigma}^2_{\theta 1}$) of their first taste shock, $\tilde{\theta}_{i1}^{pk}$. (Recall $\tilde{\sigma}^2_{\theta 1}$ is related to model parameters by equation (5).) Based on October-November 2002 pricing data (ignoring free in-network calling), Figure 6 (top panel) shows plan-choice as a function of prior beliefs $\{\tilde{\mu}_{i1}^{pk}, \tilde{\sigma}^2_{\theta 1}\}$ given $\beta = 4$ and $\varphi = 1/2$. Consumers joining in October-November 2002 with beliefs in the gray

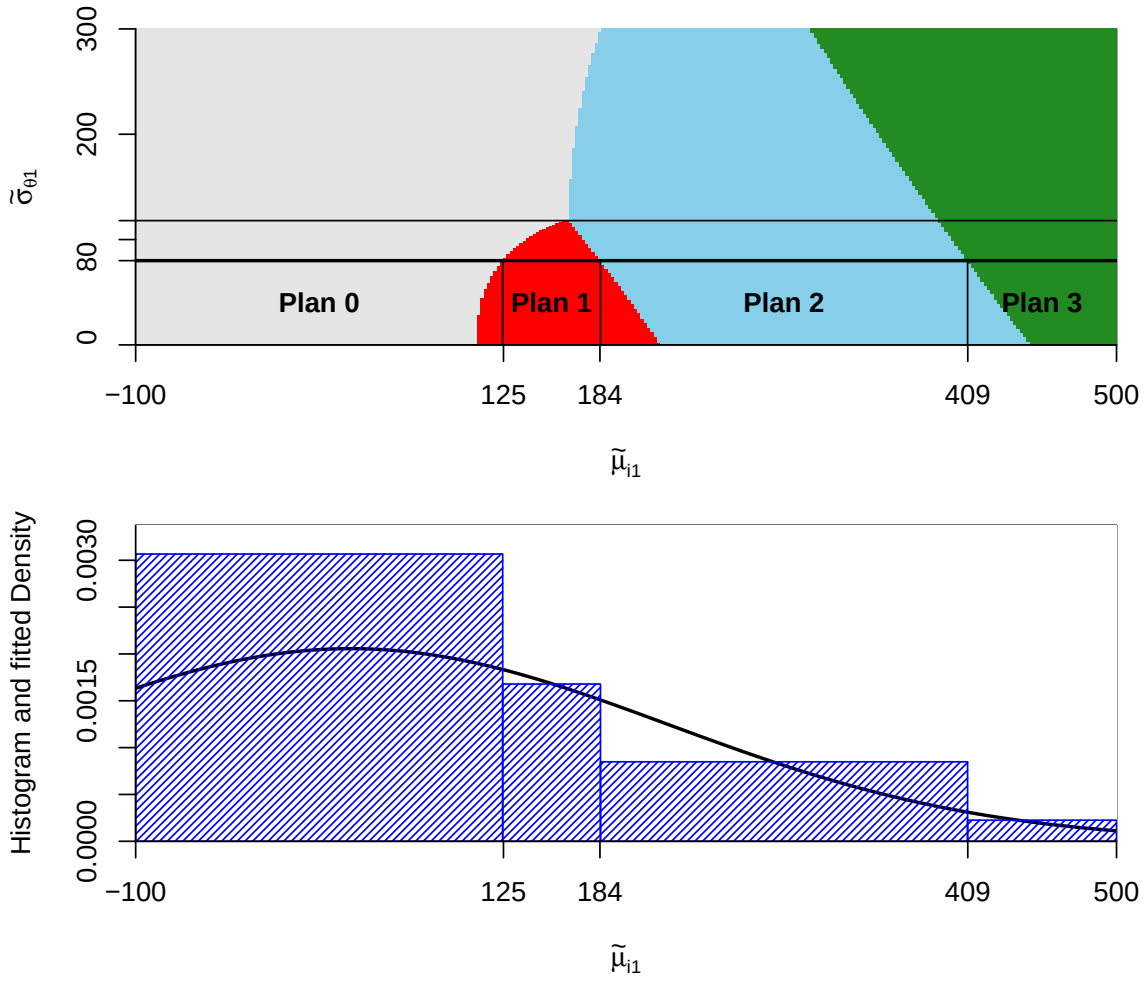


Figure 6: Top panel: Plan choice as a function of initial beliefs $\{\tilde{\mu}_{i1}, \tilde{\sigma}_{\theta 1}\}$ implied by the model evaluated at October-November 2002 prices given $\beta = 4$ and $\varphi = 1/2$. Bottom panel: Histogram and fitted normal distribution over $\tilde{\mu}_{i1}$ implied by the assumption $\tilde{\sigma}_{\theta 1} = 80$ and October-November 2002 new subscriber plan choice shares of 69%, 10%, 19%, and 2% for plans 0 to 3 respectively.

region choose plan 0, those with beliefs in the red region choose plan 1, those with beliefs in the blue region choose plan 2, and those with beliefs in the green region choose plan 3. This means that observing a new customer in October-November 2002 choose plan j will bound her beliefs to be within the relevant colored region.

Figure 6 shows that plan 0 is chosen both by individuals with low expectations of usage (low $\tilde{\mu}_{i1}^{pk}$), as it has the lowest fixed fee, and by individuals with high uncertainty about usage (high $\tilde{\sigma}_{\theta 1}$), as it never charges more than 11 cents per minute and is therefore a safe option. Figure 6 shows that for any $\tilde{\sigma}_{\theta 1}$ larger than 118, plan 1 is never chosen. Thus the assumption that $\tilde{\sigma}_{\theta 1}$ is common across individuals and the fact that a sizable fraction of individuals chose plan 1 in

October-November 2002 puts an upper bound on $\tilde{\sigma}_{\theta 1}$ of 118.

If we were to fix $\tilde{\sigma}_{\theta 1}$ at any level below 118, individual i's plan choice bounds $\tilde{\mu}_{i1}^{pk}$ to an interval. For instance, if overconfidence and volatility bias were complete ($\tilde{\sigma}_1 = \delta_\mu = \delta_\varepsilon = 0$) so that consumers believed they could predict their usage perfectly ($\tilde{\sigma}_{\theta 1} = 0$) and consumers were inelastic ($\beta = 0$), then consumers would choose from the lower envelope of the tariff menu, and initial choice of plan j would imply the following bounds on the prior point estimate $\tilde{\mu}_{i1}^{pk}$:

$$\left(M_j - M_{j-1}\right)/p_{j-1} + Q_{j-1} \leq \frac{\tilde{\mu}_{i1}^{pk}}{1-\varphi} \leq \left(M_{j+1} - M_j\right)/p_j + Q_j.$$

For $\tilde{\sigma}_{\theta 1}$ and β strictly positive, the bounds do not have an analytical solution but can be read from the corresponding horizontal slice of Figure 6. For example, the bounds are given for $\tilde{\sigma}_{\theta 1} = 80$ by the vertical lines in Figure 6. Combining plan share data from customers who join in October-November 2002 with these bounds generates of histogram over $\tilde{\mu}_{i1}^{pk}$ with four bins, one for each of the four pricing plans. Since we assume that $\tilde{\mu}_{i1}^{pk}$ is normally distributed with mean $\tilde{\mu}_0$ and standard deviation $\tilde{\sigma}_\mu$, this histogram would then (over) identify the distribution. The resulting histogram and fitted normal distribution, are both shown in the lower panel of Figure 6 for the case $\tilde{\sigma}_{\theta 1} = 80$, $\beta = 4$, and $\varphi = 1/2$.

The model identifies $\tilde{\sigma}_{\theta 1}$ as the value between 0 and 118 that generates the best fit between the histogram and the fitted normal distribution. Choosing a larger value for $\tilde{\sigma}_{\theta 1}$ implies a higher mean and a lower variance for the distribution of $\tilde{\mu}_{i1}^{pk}$.[36] Given $\beta = 4$, the overall best fit is at $\tilde{\sigma}_{\theta 1} = 83.5$.

The preceding argument for identifying $\tilde{\sigma}_{\theta 1}$, $\tilde{\mu}_0^{pk}$, and $\tilde{\sigma}_{\mu^{pk}}$ clearly bounds $\tilde{\sigma}_{\theta 1} \leq 118$ (given $\beta = 4$) but then relies heavily on the functional form assumption that $\tilde{\mu}_{i1}^{pk}$ is normally distributed for point identification. Nevertheless, there is additional information in the data which reduces reliance on the functional form assumption: As prices change over time, the bounds depicted in Figure 6 change, so that plan share data from later dates provide additional restrictions on $\tilde{\sigma}_{\theta 1}$ and the distribution of $\tilde{\mu}_{i1}^{pk}$.

The exercise described above identifies consumer uncertainty about initial tastes ($\tilde{\sigma}_{\theta 1}$) but it still remains to separate out uncertainty about own type ($\tilde{\sigma}_1$) from perceived taste volatility ($\tilde{\sigma}_\varepsilon^{pk}$), which in turn will distinguish overconfidence (δ_μ) from volatility bias (δ_ε). By equation (5), $\tilde{\sigma}_{\theta 1}^2$ is a weighted sum of $\tilde{\sigma}_1^2$ and $(\tilde{\sigma}_\varepsilon^{pk})^2$. The two parameters are distinguished by the rate of learning

[36] This is because higher uncertainty (higher $\tilde{\sigma}_{\theta 1}$) leads individuals who choose plans 1-3 to insure themselves by choosing plans with more included minutes. They are willing to choose plan 2 over plan 1 and plan 3 over plan 2 at lower values of $\tilde{\mu}_{i1}^{pk}$. However, they are only willing to choose plan 1 over plan 0 at higher values of $\tilde{\mu}_{i1}^{pk}$.

and plan switching, which is decreasing in $\tilde{\sigma}_{\varepsilon}^{pk}/\tilde{\sigma}_1$. This is apparent from the expression for $\tilde{\mu}_{i,t+1}^{pk}$ derived from Bayes rule in Appendix B.3 equation (16), which shows that a consumer's updated beliefs are a weighted average of her prior and her signals, where the weight placed on her prior is proportional to $(\tilde{\sigma}_{\varepsilon}^{pk}/\tilde{\sigma}_1)^2$. Recall that we identify the probability of an active choice based on the rate at which consumers switch away from dominated plans. Thus we can distinguish slow learning from a failure to actively consider switching.

5.3.1 Logit Error Weight

The preceding discussion ignores logit-errors, which the model does incorporate into plan choice. As a result, plan choices do not actually give sharp bounds on prior beliefs, but rather smooth likelihoods over priors, since beliefs outside the bounds described by Figure 6 can be explained by the logit error. Without logit-errors, all initial plan choices could be rationalized by prior beliefs. However, the model requires logit-errors to rationalize switches that appear to be in the 'wrong' direction. For example, suppose a customer with high average usage chooses a small plan and subsequently experiences a string of overage charges. A low prior belief ($\tilde{\mu}_{i1}^{pk}$ small) could rationalize the initial choice of a small plan. However, given the assumption of Bayesian learning, no prior can simultaneously rationalize the initial choice and a subsequent switch to an even smaller plan. The degree to which switching is in the wrong direction identifies the logit error weight $1/\alpha$.

5.4 Tastes

Having identified beliefs it is straightforward to identify taste process parameters. Given the AR1 coefficient φ, the price coefficient β, and consumer beliefs, we can calculate v_{it}^k for $k \in \{$pk-in,pk-out,op-in,op-out$\}$ and infer taste-shocks $\boldsymbol{\theta}_{it}$ and $\mathbf{r}_{it}^{9pm}$ from usage. Observing r_{it}^k for $k \in \{$9pk,9op$\}$ (a censoring of $\tilde{r}_{it}^k = \alpha_i^k + e_{it}^{r,k}$) identifies $E\left[\alpha_i^k\right]$, $Var(\alpha_i^k)$, and $Var(e_{it}^{r,k})$.[37] Correlation between observed usage and initial plan choices identifies $\boldsymbol{\psi}$, which determines the correlation between beliefs and true types. Given φ and $\boldsymbol{\theta}_{it}$, we can calculate the composite error $(\boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_{it}) = \boldsymbol{\theta}_{it} - \varphi\boldsymbol{\theta}_{i,t-1}$, which is joint-normally distributed conditional on $\tilde{\mu}_{i1}^{pk}$, so unconditionally is the mixture of joint normals. The argument for identifying this distribution is then similar to that for identifying the error structure in a random effects distribution. This delivers the parameters $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_{\mu}$, and $\boldsymbol{\Sigma}_{\varepsilon}$. Finally, bias measures δ_{μ}, δ_{ε}, b_1, and b_2 can be computed from their definitions.

[37] Without censoring, these would simply be $E\left[\alpha_i^k\right] = E\left[r_{it}^k\right]$, $Var\left(\alpha_i^k\right) = Cov(r_{it}^k, r_{it-1}^k)$, and $Var(e_{it}^{r,k}) = Var(r_{it}^k) - Var(\alpha_i^k)$.

6 Estimation Procedure

Before describing our estimation procedure, we outline the parameters to be estimated. First are those associated with beliefs: the parameters governing the distribution of consumer beliefs, $\tilde{\mu}_0^{pk}$ and $\tilde{\sigma}_{\mu^{pk}}$, consumers' initial uncertainty about their peak type, $\tilde{\sigma}_1$, and consumers' estimate of taste volatility, $\tilde{\sigma}_\varepsilon$. The parameters associated with actual tastes for usage are the means of the μ_{it}^k's, μ_0^{pk} and μ_0^{op}, their variances and correlation, $\sigma^2_{\mu^{pk}}$, $\sigma^2_{\mu^{op}}$, and ρ_μ, and the variances and correlation of the idiosyncratic errors, $(\sigma_\varepsilon^{pk})^2$, $(\sigma_\varepsilon^{op})^2$, and ρ_ε, as well as ψ^{pk} and ψ^{op}, which capture correlation between beliefs and actual usage. There are four parameters which govern the shares of outgoing landline calls occurring between 8:00 pm and 10:00 pm: the average peak share μ_α^{9pk}, the individual specific variance $(\sigma_\alpha^{9pk})^2$, and the two idiosyncratic variances $(\sigma_e^{9k})^2$ for $k \in \{pk, op\}$.[38] The final set of parameters that are discussed in the text include the price coefficient β, the logit error weight $1/\alpha$, the active choice probability P_C, and the outside good utility O. Finally, we estimate an additional six parameters that govern the share of in-network usage and a parameter that reflects consumer beliefs about the share of in-network usage. We discuss these parameters further in Appendix C. We denote the vector of all parameters as Θ, which is 30 dimensional.

We begin this section by describing the structure of the likelihood function which arises from our model. As discussed below, the likelihood function for our model does not have a closed form expression due to the presence of unobserved heterogeneity. We therefore turn to Simulated Maximum Likelihood to approximate the likelihood function (Gourieroux and Monfort 1993).

An observation in our model is a usage plan-choice pair for a consumer at a given date. At each observation, we must evaluate the joint likelihood of observed usage and plan choice conditional on observed prices and the consumer's usage and choice history. The likelihood for an observation arises naturally from the distributional assumptions on our model's unobservables. To facilitate the exposition, we divide the unobservables into two groups. The first group consists of random variables that are independent across individuals, but are not independent across time within an individual. This consists of the unobservables $\tilde{\mu}_{i1}^{pk}$, μ_i^{pk}, μ_i^{op}, α_i^{9pk}, two normally distributed individual specific effects which govern the share of in-network usage for peak and off peak, α_i^{pk} and α_i^{op}, and latent $\tilde{\theta}_{it}^k$ when $\theta_{it}^k = 0$ for $k \in \{pk, op\}$. (When category $k \in \{pk, op\}$ usage is zero, we can infer that the censored taste shock θ_{it}^k is zero but the latent taste shock $\tilde{\theta}_{it}^k \leq 0$ is unobserved.) We group these random variables together into a vector denoted $\boldsymbol{u}_i$. The second group of error terms consist of structural shocks that are independent across time and individuals: the logit plan

[38] Recall that we do not need to estimate a mean or individual specific variance for off peak 9:00 pm to 10:00 pm usage because we restrict average peak and off-peak tastes for 8:00 pm to 10:00 pm usage to be equal in equation (7).

choice error η_{itj}, the errors in the stochastic process of $\tilde{\theta}_{it}^k$ when $\tilde{\theta}_{it}^k > 0$, ε_{it}^k, idiosyncratic errors for the 8:00 pm to 10:00 pm shares, r_{it}^{9k} and e_{it}^k for $k \in \{9pk, 9op\}$, as well as two normally distributed idiosyncratic errors governing in-network usage, which we denote e_{it}^{pk} and e_{it}^{op}, respectively.

For individual i at time period t, we observe a plan choice j as well as a vector of usage, $\boldsymbol{q}_{it}$, where $\boldsymbol{q}_{it} = \{q_{it}^{pk,in}, q_{it}^{pk,out}, q_{it}^{op,in}, q_{it}^{op,out}, q_{it}^{9pk}, q_{it}^{9op}\}$, and the *in* and *out* superscripts refer to in-network and out-of-network usage. Conditional on $\mathbf{u}_i$, the likelihood of an observation will simply be the product of the choice probability and the likelihood of the observed usage.

First, consider the choice probability. Conditional on information set $\Im_{it}$ and an active choice in period t,[39] an individual will choose plan j when that plan has the highest utility according to equation (3). Let J_{it} denote the set of plans available to consumer i in period t. Conditional on an active choice, our assumption of logit errors gives rise to the following choice probability:

$$P_{it}(j'|C; \Im_{it}, \boldsymbol{u}_i, J_{it}) = \frac{\exp(U_{ijt}(\Im_{it}, \boldsymbol{u}_i))}{\sum_{k \in J_{it}} \exp(U_{ikt}(\Im_{it}, \boldsymbol{u}_i))}.$$

Unconditional on an active choice, the probabilities that an existing customer switches to plan j' in period t (where j' could be the outside good) or keeps the existing plan j are $P_C P_{it}(j'|C; \Im_{it}, \boldsymbol{u}_i, J_{it})$ and $P_C P_{it}(j|C; \Im_{it}, \boldsymbol{u}_i, J_{it}) + (1 - P_C)$ respectively:

$$P(\text{Choose } j'|\Im_{it}, \boldsymbol{u}_i, J_{it}) = \begin{cases} P_C P_{it}(j'|C; \Im_{it}, \boldsymbol{u}_i, J_{it}) & \text{if } j' \neq j \\ P_C P_{it}(j'|C; \Im_{it}, \boldsymbol{u}_i, J_{it}) + (1 - P_C) & \text{if } j' = j \end{cases}. \tag{9}$$

The consumer's information set in period t will contain some of the random draws, as well as past $\boldsymbol{q}_{it}$'s which impact the Bayesian updating process. The consumer's choice set J_{it} depends on the plan choices drawn from the non-university plans and the consumer's past plan choices. For a new customer, the initial choice set J_{i1} includes plans currently offered through the university but does not include the outside option or any other plans, and does not vary with the simulation draw. Other options are not included for new customers because we only observe consumers who sign up; hence the probability of plan choice for these customers is the probability of choosing plan j conditional on signing up. For existing customers, the choice set J_{it} also includes the customer's existing plan, those currently offered by the other provider considered, and the outside good. We assume that the consumer considers only one outside provider (AT&T, Cingular, or Verizon), in addition to the possibility of quitting each month. The option considered is drawn from a discrete distribution which assigns probability 1/3 to each of the three providers.

[39] Notation: conditioning on C means conditioning on an active choice.

As there are three possible choice sets, we index each choice set by J_{it}^k, $k = 1, ..., 3$.

Next, consider the likelihood of observed usage. A consumer's observed usage, $\boldsymbol{q}_{it}$, will be a function of $\boldsymbol{u}_i$, the idiosyncratic errors ε_{it}^k and e_{it}^k, and past values of $\boldsymbol{q}_{it}$ for $t > 1$. Conditional on $\boldsymbol{u}_i$ and $\boldsymbol{q}_{i1}, ..., \boldsymbol{q}_{i,t-1}$, the distributions of ε_{it}^k and e_{it}^k will generate a distribution for $\boldsymbol{q}_{it}$. We denote this density function as $f_{\boldsymbol{q}}(\boldsymbol{q}_{it}|\boldsymbol{u}_i, \boldsymbol{q}_{i1}, ..., \boldsymbol{q}_{i,t-1}, \Theta)$. Given assumed distributions of ε_{it}^k and e_{it}^k, we derive the distribution of $\boldsymbol{q}_{it}$ using a change of variables. We describe the exact form of $f_{\boldsymbol{q}}$ in Appendix D.

The likelihood of a sequence of observed usages and plan choices will be the product of the individual usage and choice likelihoods with the unobservable $\mathbf{u}_i$ integrated out:

$$L_i(\Theta) = \int_{\boldsymbol{u}_i} \prod_{t=1}^{T_i} \left[\left(\sum_{k=1}^{3} \frac{1}{3} P(\text{Choose } j' | \Im_{it}, \boldsymbol{u}_i, J_{it}^k) \right) f_{\boldsymbol{q}}(\boldsymbol{q}_{it}|\boldsymbol{u}_i, \boldsymbol{q}_{i1}, ..., \boldsymbol{q}_{i,t-1}) \right] f_{\boldsymbol{u}}(\boldsymbol{u}_i) d\boldsymbol{u}_i. \tag{10}$$

Because it has no closed form solution, we approximate the integral over $\boldsymbol{u}_i$ in equation (10) using Monte Carlo Simulation. For each individual, we take S draws on the random effects from $f_{\boldsymbol{u}}(\boldsymbol{u}_i)$ and the choice sets J_{it}^k, and approximate the likelihood using

$$\hat{L}_i(\Theta) = \frac{1}{S} \sum_{s=1}^{S} \left[\left(P(\text{Choose } j' | \Im_{it,s}, \boldsymbol{u}_{is}, J_{it,s}^k) \right) f_{\boldsymbol{q}}(\boldsymbol{q}_{it}|\boldsymbol{u}_{is}, \boldsymbol{q}_{i1}, ..., \boldsymbol{q}_{i,t-1}) \right].$$

The model log-likelihood is the sum of the logarithms of the individual likelihoods:

$$\hat{LL}(\Theta) = \sum_{i=1}^{I} \log(\hat{L}_i(\Theta)). \tag{11}$$

It is well-known that the value of Θ which maximizes $\hat{LL}$ is inconsistent for fixed S due to the logarithmic transformation in equation (11). However, it is consistent if $S \to \infty$ as $I \to \infty$, as discussed in Hajivassiliou and Ruud (1994). We chose $S = 300$; to arrive at this value we conducted some simple artificial data experiments where we simulated our model and attempted to recover the parameters, finding that 300 draws was sufficient to recover the true parameter draws to roughly 5% accuracy. We also found in our experiments that we were able to reduce simulation bias significantly by using a deterministic Sobol sequence generator to create the random draws, rather than canonical random number generators. Goettler and Shachar (2001) describe some of the advantages of this technique in detail. We use the algorithm provided in the R package randtoolbox to create the draws (Dutang and Savicky 2010).

A second issue that arises in the formulation of equation (11) is due to the censoring of serially correlated taste shocks. As noted above, whenever peak or off-peak usage is zero, the corresponding

latent taste shock $\tilde{\theta}_{it}^k$ is negative and unobserved. In such cases we substitute the probability that $\tilde{\theta}_{it}^k$ is censored for $f_{\boldsymbol{q}}$. However, we always need a value of $\tilde{\boldsymbol{\theta}}_{it}$ to calculate period $(t+1)$'s likelihood because we assume both that $\tilde{\boldsymbol{\theta}}_{it}$ follows an AR1 process and that consumers observe $\tilde{\boldsymbol{\theta}}_{it}$ when updating their $(t+1)$ beliefs. Hence, when censored, we also draw a value of $\tilde{\theta}_{it}^k$ using an importance-sampling procedure and include it in $\boldsymbol{u}_i$ to be integrated out. Our approach, which results in a smooth likelihood, is an adaptation of Lee's (1999) procedure for integrating out serially correlated latent unobservables in dynamic Tobit models.

Additional details about the likelihood function, including a treatment of in-network calling, are in Appendix D. We describe the computational procedures we use to evaluate and maximize the likelihood function in Section E. The computational difficulties in the estimation arise primarily from two sources: one is the high dimensional unobserved heterogeneity, which requires many evaluations of the likelihood function. The second is the computation of $\mathbf{v}^*$. Because there is no closed-form solution for $\mathbf{v}^*$, we use a nonlinear equation solver to solve for it numerically. We must do this for each simulation draw, at each time period, for every individual, at every choice that is not the outside good or the two-part tariff, plan 0.

7 Results

7.1 Parameter Estimates

Our parameter estimates are shown in Table 3. The first three columns show the coefficients, estimates, and standard errors for the first 15 parameters, while the fourth through sixth columns show the same for the next 15 parameters. The calling price coefficient β is 4.02, which indicates that a price increase from 0 cents to 11 cents per minute decreases usage by 31%.

The next 10 parameters characterize the distribution of $\tilde{\mu}_{i1}^{pk}$ as well as the perceived and true distributions of μ_i^{pk} and μ_i^{op} conditional on $\tilde{\mu}_{i1}^{pk}$. On average, consumers believe their mean draw of $\tilde{\theta}_{it}^{pk}$ to be negative 19, while the actual mean of $\tilde{\theta}_{it}^{pk}$ is 103 minutes. (Accounting for censoring of the latent shock, the average consumer believes the mean of θ_{it}^{pk} is 72 minutes and the true mean is 175 minutes.) The average off-peak draw $\tilde{\theta}_{it}^{op}$ is slightly below the peak value at 97 minutes. (The model predicts higher off-peak usage due to consumer price sensitivity.)

The standard deviation in consumers' initial belief $\tilde{\mu}_{i1}^{pk}$ is 143 minutes. Conditional on $\tilde{\mu}_{i1}^{pk}$, the standard deviation of consumer uncertainty about true type, $\tilde{\sigma}_1$, is 12.9 minutes. In contrast, conditional on $\tilde{\mu}_{i1}^{pk}$, the population standard deviations of μ_i^{pk} and μ_i^{op} are 78 and 162 minutes respectively. Thus consumers are overconfident, underestimating uncertainty about true type by 84%. The estimates of $\boldsymbol{\psi}$ indicate that initial beliefs are slightly negatively correlated with μ_i^{pk}

Table 3: Parameter Estimates

Coefficient	Estimate	Std. Err	Coefficient	Estimate	Std. Err
β	4.024	(0.02)	μ_α^{9pk}	-0.004	(0.001)
$\tilde{\mu}_0^{pk}$	-19.275	(1.33)	$(\sigma_\alpha^{9pk})^2$	0.06	(0.001)
μ_0^{pk}	103.327	(1.896)	$(\sigma_e^{9pk})^2$	0.104	(0.001)
μ_0^{op}	97.512	(3.418)	$(\sigma_e^{9op})^2$	0.116	(0.001)
$\tilde{\sigma}_{\mu^{pk}}$	143.196	(0.646)	φ	0.579	(0.008)
$\tilde{\sigma}_1$	12.876	(0.046)	α	0.096	(0.163)
$\sigma_{\mu^{pk}}$	78.584	(1.462)	Price Consideration	0.063	(0.046)
$\sigma_{\mu^{op}}$	161.784	(2.235)	Outside Good Utility	-71.191	(30.312)
ψ^{pk}	-0.043	(0.012)	δ_r	0.003	(0.037)
ψ^{op}	0.228	(0.017)	μ_α^{pk}	0.35	(0.002)
ρ_μ	0.981	(0.002)	μ_α^{op}	0.401	(0.002)
$\tilde{\sigma}_\varepsilon^{pk}$	163.574	(1.746)	$(\sigma_\alpha^{pk})^2$	0.035	(0.001)
σ_ε^{pk}	182.521	(0.452)	$(\sigma_\alpha^{op})^2$	0.039	(0.001)
ρ_ε	0.407	(0.004)	$(\sigma_e^{pk})^2$	0.03	(0)
σ_ε^{op}	306.574	(0.66)	$(\sigma_e^{op})^2$	0.025	(0)
Log-likelihood	264684.7				

and are positively correlated with μ_i^{op}. Finally, conditional on $\tilde{\mu}_{i1}^{pk}$ the correlation between peak and off-peak μ_i^k is high at 98%. The unconditional correlation between peak and off-peak μ_i^k is somewhat lower at 94%; the unconditional standard deviations of the peak and off-peak μ_i^k are slightly higher than their conditional values at 79 minutes and 165 minutes respectively.

The last four rows of column 1 describe the perceived and true distributions of the error term $\boldsymbol{\varepsilon}$. The perceived standard deviation of ε_{it}^{pk}, $\tilde{\sigma}_\varepsilon^{pk}$, is 164 minutes. In contrast, the true standard deviation is 183 minutes, meaning consumers underestimate volatility by 10%. The variances of peak and off peak errors are higher than the unconditional variances of μ_i^{pk} and μ_i^{op}, indicating that more of the variation in usage can be attributed to monthly volatility than the consumer-level fixed effect; additionally, their correlation is much lower.

The first four parameters of column 2 describe consumers' tastes for 8:00 pm to 10:00 pm usage. The low value of μ_α^{9pk} indicates that outgoing 8:00 to 9:00 pm landline usage is small as a fraction of total peak usage, which is consistent with the data. The φ value of 0.58 indicates strong serial correlation in tastes from month to month. The logit error scaling parameter, α, is estimated to be a little less than 0.1. The price consideration parameter is 0.063, indicating that consumers seldom look at prices, but it is not precisely estimated. The imprecision is consistent with our artificial data experiments, where we found that this parameter was difficult to identify. The outside good utility

is estimated to be -71. Compared to average utilities of about -20, this implies that consumers prefer inside goods to the outside good by a large margin.

The last seven parameters relate to in-network usage. We describe the modifications to the model needed to distinguish in and out-of-network usage in Appendix C. Loosely, the parameter δ_r measures consumers' underestimation of the fraction of calls that are in-network. Since our estimate of δ_r is close to zero, consumers believe that almost all usage is out of network.[40] The next two parameters govern the shares of $\boldsymbol{\theta}_{it}$ which can be apportioned to peak and off-peak in-network usage, respectively, while the final four govern the variances of in-network usage.

7.2 Biases and Learning

Returning to consumer beliefs, the parameters which summarize consumer biases are functions of our estimated parameters. We display estimates of these parameters in Table 4. Our estimates of δ_μ and δ_ε indicate strong overconfidence and mild volatility bias, respectively. Consumers underestimate their uncertainty about their own average tastes by 84% and underestimate the monthly volatility in their tastes by 10%. Together, overconfidence and volatility bias imply that the standard deviation of consumers' initial uncertainty about $\tilde{\theta}_{i1}^{pk}$, $\tilde{\sigma}_{\theta 1}$, is 203 minutes rather than the correctly calibrated 292 minutes. (Note that if consumers are risk averse rather than risk neutral then these estimates are lower bounds on the magnitudes of overconfidence and volatility bias.) Aggregate mean bias is negative, indicating that the average consumer underestimates her initial $\tilde{\theta}_{it}^{pk}$ draw by 123 minutes. Finally, the positive estimate of b_2 reflects strong positive conditional mean bias.[41]

Table 4: Estimates of Consumer Beliefs

Coefficient	Estimate	Std. Err
δ_μ	0.164	(0.003)
δ_ε	0.896	(0.01)
b_1	-122.602	(2.19)
b_2	1.043	(0.012)

The fact that overconfidence is stronger than volatility bias ($\delta_\mu < \delta_\varepsilon$) implies that consumers

[40] Plan 0 always offered free in-network usage and plan 2 did so as well near the end of our sample period. We incorporated this parameter to help explain the high share of Plan 1 relative to Plan 0, as plan 0 dominates plan 1 for anyone with a median in-network usage share.

[41] In the context of grocery home delivery service, Goettler and Clay (2011) also find $b_1 < 0$ and $b_2 > 0$.

overweight their priors relative to new experience and hence learn slowly. This is illustrated in Figure 7, which plots an average of consumers' evolving point-estimates $\tilde{\mu}_{it}^{pk}$ for consumers whose true value is $\mu_i^{pk} = \mu_0^{pk} \approx 103$. A consumer's time t point-estimate $\tilde{\mu}_{it}^{pk}$ is a function of her initial belief $\tilde{\mu}_{i1}^{pk}$ and her past taste shocks $\boldsymbol{\theta}_i^{t-1}$. The dotted lines in the figure show the average of $\tilde{\mu}_{it}^{pk}$ for 1000 simulated consumers, where each consumer's $\tilde{\mu}_{i1}^{pk}$ and $\boldsymbol{\theta}_i^T$ are drawn from their estimated distributions. The thick red line shows how consumers' beliefs evolve given estimated overconfidence and volatility bias. An average consumer whose true μ_i^{pk} is roughly 103 minutes and who enters the sample believing $\tilde{\mu}_{i1} = \tilde{\mu}_0 = -19$ increases her belief to $\tilde{\mu}_{i,13} = -9.6$ after one year (an 8% reduction in aggregate mean bias). The blue dashed line shows how beliefs evolve when overconfidence and volatility bias are removed. Debiasing consumers speeds up learning and after 1 year an average consumer's belief about μ_i^{pk} will be $\tilde{\mu}_{i,13} = 66$ (a 70% reduction in aggregate mean bias).

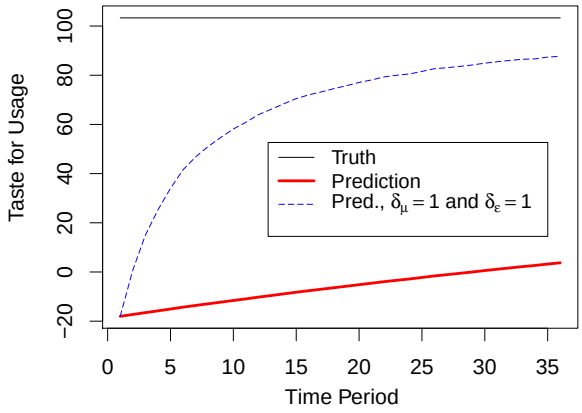


Figure 7: Population average $\tilde{\mu}_{it}^{pk}$ (point estimate of μ_i^{pk}) for those with true value $\mu_i^{pk} = \mu_0^{pk} \approx 103$.

7.3 Fixed-Price Counterfactual: Impact of Biased Beliefs

Before proceeding to simulate endogenous price changes in Section 8, we briefly simulate the change in firm profits, consumer welfare, and total welfare that results from debiasing consumers while holding observed prices fixed (Table 5). We construct these counterfactual simulations at our data in the sense that we hold fixed the number of consumers, and when consumers enter and exit the data set. Surplus changes are measured in dollars per student over the two year period that they are observed and assume marginal cost is zero. The first three columns of Table 5 show the welfare effects when students face university prices, while the last three columns show the welfare effects when consumers face publicly available prices. The consequences of debiasing are larger in the latter case because the university's plan 0, which is absent from the public menu, tended to protect biased consumers.

Table 5: Counterfactual: Per student change in surpluses from bias elimination (fixed prices)

	University Plans			Public Plans		
Beliefs	Profits	Cons. Welf.	Total Welf.	Profits	Cons. Welf.	Total Welf.
$\delta_\mu = 1$	−23.09	16.05	−7.04	−48.91	33.09	−15.82
$\delta_\varepsilon = 1$	−6.59	4.86	−1.73	−12.41	9.07	−3.34
$\delta_\mu = 1$ and $\delta_\varepsilon = 1$	−23.56	15.21	−8.35	−51.88	34.44	−17.44
No Biases	−44.01	36.95	−7.07	−58.54	42.21	−16.34

Changes in surpluses (profits, consumer welfare, and total welfare) are measured in dollars per student over the 2 year sample period. Changes are relative to surpluses at estimates.

The first row of Table 5 shows the impact of removing overconfidence, which raises consumer surplus but lowers profits and total welfare. Row two shows similar effects of removing volatility bias and row three shows the combined effects of removing both biases, which raises consumer surplus by $34 given public prices. Finally, row four shows the total effect of removing all biases, including mean biases and underestimation of in-network calling (discussed in Appendix C), which raises consumer surplus by $42 given public prices. On average, debiased consumers are less likely to choose plan 1 and make fewer calls because they are more aware of overage risk. The reduction in plan 1 share is more pronounced for public prices because its initial share is higher without plan 0. The reduction in calling reduces total welfare because marginal costs are approximately zero. Thus gains in consumer surplus are overshadowed by profit losses.

8 Endogenous-Price Counterfactual: Bill-Shock Regulation

8.1 Nested-Logit Specification

To predict the effect of bill-shock regulation on equilibrium prices it is important to correctly capture the degree of competition between carriers. Hence, we modify the error structure of the demand model to be a two level nested logit, rather than logit. In our nested-logit specification, we assume that each inside nest contains the plans offered by a carrier (the option of shutting off cellular-phone service is also put in its own nest). The outside nest consists of all the carriers (including no service) in a particular consumer's consideration set.[42] We assume that the inclusive value parameter, denote by λ, is the same for each option.

The more restrictive logit specification implies that if consumers choose plans within carrier

[42] A new consumer chooses among three carriers and no-serivce whereas an existing consumer who considers switching chooses among her current carrier, a randomly chosen outside carrier, and no service.

primarily based on price then carriers are close substitutes. Thus the logit specification leads to unrealistically high competition and low prices in counter-factual simulations. We choose the more flexible nested-logit specification because it allows consumers to have strong idiosyncratic carrier preferences (due to network coverage or phone availability) that create market power, while at the same time making within carrier plan choices primarily based on price.

Ideally, we would like to estimate λ jointly with the other demand parameters using demand side choice data. Unfortunately, we observe neither carrier market shares on campus nor the alternate carriers chosen by students quitting university plans. Hence only the quitting rate is available to identify utility of the outside good, average utility of university plans relative to other carriers, and λ. In our demand estimates we assume $\lambda = 1$ (logit specification) and carrier symmetry to identify the outside good utility.[43] To address this identification problem, we calibrate λ using supply-side price data: We select the value of λ that best rationalizes observed prices conditional on our demand estimates. Our algorithm, which is described in Appendix F, calibrates λ to be 0.2.

Before proceeding, we make two comments on our calibration approach. First, one potential problem is that our demand estimates were made conditional on $\lambda = 1$ (which generates the logit model), but different values of λ might produce different demand estimates. Fortunately, our demand estimates are relatively insensitive to λ, which we show in Appendix F. Second, in principle we could have estimated λ and the other parameters jointly by using constrained maximum likelihood and constraining observed prices to be optimal at the estimated parameters. We avoided this approach because we prefer only to impose our supply-side structural assumptions (that competition is symmetric static Nash in prices and that our student population is representative[44]) only when they are necessary in the endogenous-price counterfactual simulations.

8.2 Modeling bill-shock alerts and endogenous prices.

In our bill-shock regulation counterfactual, consumers are informed when their usage reaches Q, their allotment of free minutes.[45] In response to this new policy, a consumer's usage rule changes: A consumer will accept all calls valued above v^* until she exhausts her included minutes. After that

[43] Outside price variation is too limited to separately identify λ. Moroever, in an unreported specification, we instead chose a natural normalization for the outside good and estimated λ. We rejected this alternative, however, because the resulting estimate of λ was zero, an implausible number that implies carriers have monopoly power. This may have been due to the fact that our normalization of the outside good was too low, that carrier symmetry is a bad assumption when including the university plans, or the fact that forced quits due to graduations (outside the model) biased the estimate downwards.

[44] In reality, university plans are not symmetric to other carrier offerings and our population of students is likely overweighted towards new and low-volume users relative to the overall population.

[45] Alerts are not applicable to two-part tariffs with constant marginal prices.

point, she only accepts calls valued above p. Because the consumer adjusts her calling threshold upon making Q calls, the optimal initial threshold v^* differs from that characterized by equation (2). Appendix B.2 describes expected utility and characterizes v^* under bill-shock regulation. To calculate endogenous equilibrium prices, we assume that there are three symmetric carriers, equilibrium is symmetric static Nash in prices, marginal costs are zero, overage rates are at most fifty cents,[46] and each carrier offers a menu of three plans.

8.3 Counterfactual Simulation Results

Table 6 shows the results of our endogenous-price counterfactual simulations. Column 1 shows predicted plan prices and welfare outcomes under our estimated demand parameters. The model predicts that firms offer a two-part tariff at fifty cents per minute for $26.70 per month, a three-part tariff with 349 included minutes for $60.17 per month, and an unlimited plan for $75.58 per month.

Table 6: The Impact of Bill Shock Regulation and Removing Biases on Equilibrium Prices

			Est, Bill Shock		$\delta_\mu = 1$	
		Est	(fixed prices)	Est, Bill Shock	and $\delta_\varepsilon = 1$	No Biases
		(1)	(2)	(3)	(4)	(5)
Plan 1	M	26.70	26.70	26.42	26.15	74.43
	Q	0	0	0	0	∞
	p	50	50	50	50	N/A
	Share	54	53	56	48	33
Plan 2	M	60.17	60.17	61.73	56.27	74.43
	Q	349	349	211	0	∞
	p	50	50	12	8	N/A
	Share	28	29	26	27	33
Plan 3	M	75.58	75.58	76.52	77.15	74.43
	Q	∞	∞	∞	∞	∞
	p	N/A	N/A	N/A	N/A	N/A
	Share	18	17	19	25	33
Outside Good Share		0	0	0	0	0
Δ Annual Profit			-16.47	-0.2	8.18	32.44
Δ Annual Cons Welfare			8.34	-2.01	20.28	168.23
Δ Annual Total Welfare			-8.13	-2.21	28.47	200.67

All welfare and profit numbers are expressed in thousands of dollars. Because the counterfactuals in columns 4 and 5 produced two part tariffs, under bill shock regulation equilibrium prices are unchanged. We simulate 1000 consumers for 12 months.

[46] Otherwise the combination of biased beliefs and inattention lead to implausibly high overage-rate predictions.

Column 2 of Table 6 holds constant the predicted prices from column 1 but imposes bill-shock regulation. Holding prices constant, bill-shock alerts help plan 2 customers avoid an average of $54.97 in overage payments annually (reducing profits by the same amount). Avoided overage charges correspond to reduced calling, so average consumer surplus rises by only $27.00 per year for an average plan 2 customer. Because marginal costs are zero, the difference, $27.97, reflects the annual decrease in total welfare per plan 2 customer due to reduced calling.[47] (Column 2 of Table 6 reports these figures averaged across all customers.)

Column 3 of Table 6 imposes bill-shock regulation but allows firms to adjust prices. In equilibrium, markups are determined primarily by the calibrated inclusive value parameter, which at $\lambda = 0.2$ implies markups of about $74 per month. As a result, following bill-shock regulation, firms adjust the price of plan 2 to compensate for lost overage revenue and maintain a stable markup. Hence annual profits are stable (falling by only $0.20 per consumer per year) and consumers lose (an average of $2.01 per consumer per year) because they essentially become residual claimants of total welfare.[48]

As before, plan 2 customers are most affected by bill-shock regulation. Now, however, there are two distinct groups to consider – those who continue to choose plan 2 and those who switch to plan 1 in response to price changes. Those who continue to choose plan 2 after bill-shock regulation is implemented only benefit by an average of $15.53 per year due to changes in plan 2 pricing, including the $1.56 monthly fee increase and the 138 minute included allowance reduction. However, their contribution to total welfare actually increases because they make more calls in response to the lower overage rate of 12 cents per minute.

The preceding good news, that bill-shock alerts raise consumer surplus and improve efficiency for those who continue to choose plan 2, is only part of the story. Although average consumer surplus rises by $15.53 on plan 2, consumers do not appreciate this and actually perceive a $31.76 drop in annual surplus. This is because much of the benefit comes from the ability to respond to bill-shock alerts and take advantage of a low overage rate in months with high usage. Unfortunately, biased consumers underestimate the incidence of such high usage and therefore underestimate both the value of bill-shock alerts the value of a lower overage rate. In contrast, even biased consumers fully appreciate the cost of a $1.56 monthly fee increase. Thus, on the margin, more consumers

[47] Holding average total calling constant, bill-shock alerts reduce welfare by inducing consumers to call more in low demand months (by choosing a lower v^*) and to call less in high demand months (when receiving an alert). This reduces the average value of placed calls. In addition, average total calling is reduced because biased consumers choose too low a calling threshold v^*, but correct their behavior following a bill-shock alert.

[48] An additional counterfactual simulation shows that a single firm which introduced bill-shock alerts on its own would lose $3.69 per customer annually.

choose plan 1 over plan 2. Moreover, this is a bad decision for those marginal consumers because they talk more than they anticipate. As a result, average consumer surplus falls by $309.60 annually for those who switch to plan 1. Moreover, their contribution to total welfare falls as well because they substantially reduce calling in response to the 50 cent per minute rate.

In sum, we find that bill-shock alerts are neutral or beneficial for most consumers but severely harm a minority. Averaging across all consumers yields the negative results in column 3: average annual losses of $2.21 to social welfare and $2.01 to consumer surplus due to bill-shock regulation.

Turning to columns 4 and 5 of Table 6, we investigate the consequences of debiasing consumers. Column 4 shows the effect of eliminating overconfidence and projection bias. In column 4, plans 1 and 3 are similar to those in column 1 but plan 2 becomes a two-part tariff charging $56 per month and 8 cents per minute. Moreover, consumers tend to choose larger contracts and total welfare increases by $28 per consumer annually. Column 5 shows the effect of eliminating all biases. In this case, firms offer three identical plans that each charge $74 for unlimited calling and total welfare increases by $201 per consumer annually. In both columns 4 and 5, debiasing increases firm profits but most of the increase in total surplus accrues to consumers.

To understand these pricing results, we consider scenarios in reverse order, beginning with column 5. The unlimited calling plans in column 5 achieve first-best surplus via marginal cost pricing and earn a markup of $74 from each consumer. (Three identical such contracts do better than a single contract due to the red-bus/blue-bus problem.) When consumers are unbiased this is optimal because: (1) we assume that the inclusive value parameter is the same for all consumers, and (2) we estimate a low outside good value that predicts full market coverage. Thus, firms have no incentive to price discriminate and charge unbiased consumers different markups.

In column 4, positive conditional mean bias implies that consumers choose overly extreme plans. Thus consumers who choose plan 3 overestimate their usage while those who choose plan 1 underestimate their usage. (Negative aggregate mean bias means plan 1 consumers outnumber plan 3 consumers.) Plan 3 is optimal for over-estimators because it sells calls to consumers up front and then offers no refunds. Plan 1 is optimal for under-estimators because it extracts value ex post when consumers realize their true value for calls. Plan 2 caters to those consumers in between whose under or over estimation is mild.

Returning to columns 1-3, plan 1 and plan 3 pricing is driven by conditional mean bias following the same logic as in column 4. However, in columns 1-3 overconfidence and volatility bias predominate for intermediate plan 2 consumers and hence a three-part tariff is optimal (Grubb 2009).

Because three-part tariff pricing is driven by overconfidence and volatility bias, eliminating these biases also eliminates three-part tariff pricing. Thus bill-shock regulation has no effect without bias.

Moreover, the fact that bill-shock regulation is less important for unbiased consumers does not depend entirely on the elimination of three-part tariffs. In a final counterfactual, we simulate the effect of bill-shock regulation while holding observed public prices constant. In this simulation, bill-shock regulation benefits consumers with estimated biases by $22 but benefits debiased consumers by only $6. Debiased consumers are affected less by bill-shock alerts because (even holding prices constant) they make better plan choices that lead to lower incidence of overages.

9 Conclusion

We specify and estimate a model of consumer cellular-phone plan and usage choices. We identify the distribution of consumer tastes from observed usage and consumers' beliefs about their future usage from observed plan choices. Comparing the two we find that consumers underestimate their average taste for calling, underestimate their own uncertainty about their average tastes, and underestimate the volatility of their tastes from month-to-month. Because the magnitude of overconfidence is substantially larger than that of volatility bias, consumers correct initial plan choice mistakes more slowly than would unbiased consumers.

We conduct counterfactual simulations in which we (a) eliminate biases and (b) quantify the welfare impact of bill-shock regulation. We find that eliminating biases significantly increases consumer welfare, by $42 annually per consumer holding observed public prices fixed, and $201 annually per consumer accounting for firms' endogenous pricing response. If observed prices do not respond to bill-shock regulation, then the average consumer will benefit by $22 annually. This finding is reversed when firms optimally respond to bill-shock regulation. Although consumers avoid overage fees, firms raise monthly fees and average consumer surplus falls by $2 annually. In either case, bill-shock regulation lowers total welfare. Finally, we find that bill-shock regulation would have little to no effect if consumers were unbiased.

Our evaluation of bill-shock regulation could be insightful in other relevant contexts as well. For instance, in 2009 US checking overdraft fees totalled more than $38 billion and have been the subject of new Federal Reserve Board regulation (Martin 2010, Federal Reserve Board 2009). Convincing evidence of consumer inattention (Stango and Zinman 2009, Stango and Zinman 2010) suggests that this fee revenue would be dramatically curtailed if the Fed imposed its own bill-shock regulation by requiring debit card processing terminals to ask users "$35 overdraft fee applies, continue Yes/No?" before charging fees. Our counterfactual shows that in the cellular context consumers are nevertheless made worse off after accounting for endogenously higher fixed fees.

References

Ackerberg, Daniel A., "Advertising, Learning, and Consumer Choice in Experience Good Markets: An Empirical Examination," *International Economic Review*, 2003, *44* (3), 1007–1040.

Altschul, Michael F., Christopher Guttman-McCabe, and Brian M. Josef, "Comments of CTIA - The Wireless Association," January 10th 2011. `http://fjallfoss.fcc.gov/ecfs/document/view?id=7021025497`.

Armstrong, Mark and John Vickers, "Competitive Price Discrimination," *RAND Journal of Economics*, 2001, *32* (4), 579–605.

Ascarza, Eva, Anja Lambrecht, and Naufel J. Vilcassim, "When Talk is 'Free': The Effect of Tariff Structure on Usage under Two- and Three-Part Tariffs," *SSRN eLibrary*, 2012.

Ater, Itai and Vardit Landsman, "The Role of Overage Payments in Shaping Customers' Choice Among Three-Part Tariff Plans," 2011.

Bar-Gill, Oren and Rebecca Stone, "Pricing Misperception: Explaining Pricing Structure in the Cellular Service Market," *SSRN eLibrary*, 2009. `http://ssrn.com/paper=1425046`.

Baron, David P. and David Besanko, "Regulation and Information in a Continuing Relationship," *Information Economics and Policy*, 1984, *1* (3), 267–302.

Bhargava, Saurabh and Vikram Pathania, "Driving Under the (Cellular) Influence," Working Paper August 27 2011.

Blundell, Richard and Stephen Bond, "Initial conditions and moment restrictions in dynamic panel data models," *Journal of Econometrics*, 1998, *87* (1), 115–143.

Borenstein, Severin, "To What Electricity Price Do Consumers Respond? Residential Demand Elasticity Under Increasing-Block Pricing," Preliminary Draft April 30 2009.

Busse, M. R., "Multimarket Contact and Price Coordination in the Cellular Telephone Industry," *Journal of Economics and Management Strategy*, 2000, *9* (3), 287–320.

Cardon, James H. and Igal Hendel, "Asymmetric Information in Health Insurance: Evidence from the National Medical Expenditure Survey," *The RAND Journal of Economics*, 2001, *32* (3), 408–427.

Ching, Andrew, Tülin Erdem, and Michael P. Keane, "The Price Consideration Model of Brand Choice," *Journal of Applied Econometrics*, 2009, *24* (3), 393–420.

Chintagunta, Pradeep, Puneet Manchanda, and S. Sriram, "Empirical Investigation of Consumer Adoption, Consumption, and Termination of a Video on Demand Service," Work in Progress 2009.

Conlin, Michael, Ted O'Donoghue, and Timothy J. Vogelsang, "Projection Bias in Catalog Orders," *American Economic Review*, 2007, *97* (4), 1217–1249.

Courty, Pascal and Hao Li, "Sequential Screening," *The Review of Economic Studies*, 2000, *67* (4), 697–717.

Crawford, Gregory S. and Matthew Shum, "Uncertainty and Learning in Pharmaceutical Demand," *Econometrica*, 2005, *73* (4), 1137–1173.

CTIA - The Wireless Association, "CTIA-The Wireless Association, Federal Communications Commission and Consumers Union Announce Free Alerts to Help Consumers Avoid Unexpected Overage Charges," October 17 2011. http://www.ctia.org/media/press/body.cfm/prid/2137.

___ , "Year-End 2010 Top-Line Survey Results," Technical Report 2011. http://files.ctia.org/pdf/CTIA_Survey_Year_End_2010_Graphics.pdf.

DeGroot, Morris H., *Optimal Statistical Decisions*, New York: McGraw-Hill, 1970.

DellaVigna, Stefano and Ulrike Malmendier, "Contract Design and Self-Control: Theory and Evidence," *The Quarterly Journal of Economics*, 2004, *119* (2), 353–402.

Deloney, Amalia, Linda Sherry, Susan Grant, Parul P. Desai, Chris M. Riley, Matthew F. Wood, John D. Breyault, Jessica J. Gonzalez, and Benjamin Lennett, "Comments of the Center for Media Justice, Consumer Action, Consumer Federation of America, Consumers Union, Free Press, Media Access Project, National Consumers League, National Hispanic Media Coalition and New America Foundation Open Technology Initiative in response to notice of proposed rulemaking.," January 10th 2011. http://fjallfoss.fcc.gov/ecfs/document/view?id=7021025418.

Dutang, Christophe and Petr Savicky, "randtoolbox: Generating and Testing Random Numbers.," 2010. R package version 1.10 http://cran.r-project.org/web/packages/randtoolbox/.

Einav, Liran, Amy Finkelstein, Iuliana Pascu, and Mark Cullen, "How General Are Risk Preferences? Choices under Uncertainty in Different Domains," *American Economic Review*, Forthcoming.

Eliaz, Kfir and Ran Spiegler, "Contracting with Diversely Naive Agents," *The Review of Economic Studies*, 2006, *73* (3), 689–714.

___ **and** ___ , "Consumer Optimism and Price Discrimination," *Theoretical Economics*, 2008, *3* (4), 459–497.

Erdem, Tülin and Michael P. Keane, "Decision-Making under Uncertainty: Capturing Dynamic Brand Choice Processes in Turbulent Consumer Goods Markets," *Marketing Science*, 1996, *15* (1), 1–20.

Federal Reserve Board, "Federal Reserve announces final rules prohibiting institutions from charging fees for overdrafts on ATM and one-time debit card transactions," Press Release November 12 2009.

fei Lee, Lung, "Estimation of dynamic and ARCH Tobit models," *Journal of Econometrics*, 1999, *92* (2), 355–390.

Gaynor, Martin S., Yunfeng Shi, Rahul Telang, and William B. Vogt, "Cell Phone Demand and Consumer Learning - An Empirical Analysis," *SSRN eLibrary*, 2005.

Goettler, Ronald L. and Karen B. Clay, "Tariff Choice with Consumer Learning and Switching Costs," *Journal of Marketing Research*, 2011, *48* (4), 633–652.

____ and Ron Shachar, "Spatial Competition in the Network Television Industry," *The RAND Journal of Economics*, 2001, *32* (4), 624–656.

Gourieroux, Christian and Alain Monfort, "Simulation-based inference : A survey with special reference to panel data models," *Journal of Econometrics*, 1993, *59* (1-2), 5–33.

Grubb, Michael D., "Selling to Overconfident Consumers," *American Economic Review*, 2009, *99* (5), 1770–1807.

____ , "Bill Shock: Inattention and Price-Posting Regulation," mimeo 2011.

Hajivassiliou, Vassilis A. and Paul A. Ruud, "Classical Estimation Methods for LDV Models Using Simulation," in R.F. Engle and Daniel L. McFadden, eds., *Handbook of Econometrics, Volume IV*, Amsterdam: North-Holland, 1994.

Handel, Benjamin, "Adverse Selection and Switching Costs in Health Insurance Markets: When Nudging Hurts," 2011.

Herweg, Fabian and Konrad Mierendorff, "Uncertain Demand, Consumer Loss Aversion, and Flat-Rate Tariffs," *Journal of the European Economic Association*, Forthcoming.

Huang, Ching-I., "Estimating Demand for Cellular Phone Service Under Nonlinear Pricing," *Quantitative Marketing and Economics*, 2008, *6* (4), 371–413.

Ito, Koichiro, "Do Consumers Respond to Marginal or Average Price? Evidence from Nonlinear Electricity Pricing," 2010.

Iyengar, Raghuram, Asim Ansari, and Sunil Gupta, "A Model of Consumer Learning for Service Quality and Usage," *Journal of Marketing Research*, 2007, *44* (4), 529–544.

____ , Kamel Jedidi, and Rajeev Kohli, "A Conjoint Approach to Multipart Pricing," *Journal of Marketing Research*, 2008, *45* (2), 195–210.

Jiang, Lai, "The Welfare Effects of ''Bill Shock" Regulation in Mobile Telecommunication Markets," Working Paper July 2011.

Kim, Jiyoung, "Consumers' Dynamic Switching Decisions in the Cellular Service Industry," Working Paper, SSRN November 2006.

Lambrecht, Anja and Bernd Skiera, "Paying Too Much and Being Happy About It: Existence, Causes, and Consequences of Tariff-Choice Biases," *Journal of Marketing Research*, 2006, *43* (2), 212–223.

____ , Katja Seim, and Bernd Skiera, "Does Uncertainty Matter? Consumer Behavior under Three-Part Tariffs," *Marketing Science*, 2007, *26* (5), 698–710.

Leider, Steve and Özge Şahin, "Contracts, Biases and Consumption of Access Services," Ross School of Business Paper July 14 2011. `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1895468`.

Lichtenstein, Sarah, Baruch Fischhoff, and Lawrence D. Phillips, "Calibration of Probabilities: The State of the Art to 1980," in Daniel Kahneman, Paul Slovic, and Amos Tversky, eds., *Judgment under uncertainty : heuristics and biases*, Cambridge ; New York: Cambridge University Press, 1982, pp. 306–334.

Liebman, Jeffrey B. and Richard Zeckhauser, "Schmeduling," Working Paper October 2004.

Loewenstein, George, Ted O'Donoghue, and Matthew Rabin, "Projection Bias in Predicting Future Utility," *The Quarterly Journal of Economics*, 2003, *118* (4), 1209–1248.

Martin, Andrew, "Bank of America to End Debit Overdraft Fees," *The New York Times*, March 10 2010.

Miravete, Eugenio J., "Screening Consumers Through Alternative Pricing Mechanisms," *Journal of Regulatory Economics*, 1996, *9* (2), 111–132.

___ , "Estimating Demand for Local Telephone Service with Asymmetric Information and Optional Calling Plans," *The Review of Economic Studies*, 2002, *69* (4), 943–971.

___ , "Choosing the wrong calling plan? Ignorance and learning," *American Economic Review*, 2003, *93* (1), 297–310.

___ , "The Welfare Performance of Sequential Pricing Mechanisms," *International Economic Review*, 2005, *46* (4), 1321–1360.

___ **and Ignacio Palacios-Huerta**, "Inattention, Choice Dependence, and Learning from Experience in a Repeated Decision Problem," mimeo 2011.

___ **and Lars-Hendrik Röller**, "Estimating Price - Cost Markups Under Nonlinear Pricing Competition," *Journal of the European Economic Association*, 2004, *2* (2-3), 526–535.

Narayanan, Sridhar, Pradeep K. Chintagunta, and Eugenio J. Miravete, "The role of self selection, usage uncertainty and learning in the demand for local telephone service," *Quantitative Marketing and Economics*, 2007, *5* (1), 1–34.

Osborne, Matthew, "Consumer Learning, Switching Costs and Heterogeneity: A Structural Examination," *Quantitative Marketing and Economics*, 2011, *9* (1), 25–70.

Park, Minjung, "The Economic Impact of Wireless Number Portability," 2009.

Reiss, Peter C. and Matthew W. White, "Household Electricity Demand, Revisited," *The Review of Economic Studies*, 2005, *72* (3), 853–883.

Riordan, Michael H. and David E. M. Sappington, "Awarding Monopoly Franchises," *American Economic Review*, 1987, *77* (3), 375–387.

Rochet, Jean-Charles and Lars A. Stole, "Nonlinear Pricing with Random Participation," *The Review of Economic Studies*, 2002, *69* (1), 277–311.

___ **and Lars Stole**, "The Economics of Multidimensional Screening," in M. Dewatripont, Lars Peter Hansen, and Stephen J. Turnovsky, eds., *Advances in economics and econometrics: theory and applications - eighth world Congress*, Vol. 36 of Econometric Society Monographs, New York: Cambridge University Press, 2003.

Roodman, David, "How to do xtabond2: An introduction to difference and system GMM in Stata," *Stata Journal*, 2009, *9* (1), 86–136.

Saez, Emmanuel, "Do Taxpayers Bunch at Kink Points?," Working Paper June 2002.

___ , "Do Taxpayers Bunch at Kink Points?," *American Economic Journal: Economic Policy*, 2010, *2* (3), 180–212.

Seim, Katja and V. Brian Viard, "The Effect of Market Structure on Cellular Technology Adoption and Pricing," *American Economic Journal: Microeconomics*, 2010.

Spiegler, Ran, *Bounded Rationality and Industrial Organization*, Oxford University Press, 2011.

Stango, Victor and Jonathan Zinman, "What do Consumers Really Pay on Their Checking and Credit Card Accounts? Explicit, Implicit, and Avoidable Costs," *American Economic Review Papers and Proceedings*, 2009, *99* (2).

___ **and** ___ , "Limited and Varying Consumer Attention: Evidence from Shocks to the Salience of Overdraft Fees," 2010.

Stole, Lars A., "Price Discrimination and Competition," in Mark Armstrong and Robert K. Porter, eds., *Handbook of Industrial Organization*, Vol. Volume 3, Elsevier, 2007, chapter 34, pp. 2221–2299.

Train, Kenneth, *Discrete Choice Methods with Simulation*, 2nd ed., Cambridge University Press, 2009.

U.S. Census Bureau, "Table 1. Preliminary Annual Estimates of the Resident Population for the United States, Regions, States, and Puerto Rico: April 1, 2000 to July 1, 2010," Technical Report NST-PEST2010-01 February 2011. `http://www.census.gov/popest/states/tables/NST-PEST2010-01.xls`.

Uthemann, Andreas, "Competitive Screening of Customers with Non-Common Priors," 2005. `http://www.ucl.ac.uk/~uctpaut/research/uthemann_ncp.pdf`.

Yao, Song, Carl F. Mela, Jeongwen Chiang, and Yuxin Chen, "Determining Consumers' Discount Rates with Field Studies," 2011.

Fettered Consumers and Sophisticated Firms: Evidence from Mexico's Privatized Social Security Market

Fabian Duarte
RAND and University of Chile
fabian.duarte@RAND.org

Justine Hastings*
Brown University and NBER
justine_hastings@brown.edu

DRAFT

July 01, 2012

*Corresponding Author. We thank Brigitte Madrian, James Choi, David Cutler, Ali Hortaçsu, James Poterba, Chad Syverson and seminar participants at Brown University, Harvard, MIT, NBER, and the Yale Junior Applied Micro Lunch, for helpful comments and suggestions. Noele Aabye, Lydia Ashton, Sarah Johnston, Carolina Orellana and Unika Shrestha provided excellent research assistance. We also thank the excellent staff and leadership at CONSAR for making this project possible. Hastings gratefully acknowledges support from the National Institute of Aging, the U.S. Social Security Administration (SSA) as part of the NBER Retirement Research Consortium (RRC), Brown University and the Yale University Institution for Social and Policy Studies. The findings and conclusions expressed are solely those of the author and do not represent the views of SSA, any agency of the Federal Government or the RRC.

1. Introduction

There is growing empirical evidence that consumers may not choose optimally when faced with difficult or complex choices involving uncertainty, imperfect information, or delayed payoffs over long time horizons (Thaler and Sunstein 2008). In these situations, people may follow the path of least resistance by making decisions based on shortcuts, approximations, or readily available information as a proxy for costly optimization (Ellison 2006). For example, people may be overly sensitive to default rules or use simple heuristics when allocating resources across investments (Benartzi and Thaler 2001; Madrian and Shea 2001; Cronqvist and Thaler 2004; Choi, Laibson and Madrian 2006; Bernartzi and Thaler 2007; Beshears et al. 2008). They respond to advertising, brand name, peer opinion and irrelevant information, or focus on easy-to-understand or salient prices when making decisions (Ausubel 1991; Duflo and Saez 2003; Liebman and Zechauser 2004; McFadden 2006; Cronqvist 2006; Choi, Laibson and Madrian 2007; Kling et al. 2008; Mullainathan and Schwartzstein 2008; Chetty, Looney and Kroft 2008; Abaluck and Gruber 2009).

This implies that people may not be sufficiently adept decision-makers to incentivize efficient markets, but also suggest that government can move markets towards efficient outcomes by designing policies that facilitate informed consumer choice (McFadden 2006; Thaler and Sunstein 2008). This paper brings new evidence from the privatized social security system in Mexico, offering insight into investment behavior and the efficacy of government "nudges" in the context of profit maximizing firms.

Mexico privatized its social security system in 1997, moving from a pay-as-you-go system to a defined contribution system with individual private accounts managed by approved private fund managers. Social security and payroll taxes totaling 6.5% of salary are automatically deducted from payroll each month and placed in the personal social security (SAR) account. Workers choose between any of the approved fund managers regardless of place of employment, and between ten and twenty-one well-known firms have competed in the market since the system's inception. The reform was intended to increase equality, efficiency, and wealth at retirement through privatization of pension accounts.

Despite the large number of firms, tight investment regulations and centralized data processing, high fees persisted since the inception of the system (Hastings, Hortaçsu and Syverson (2012)). During our sample period, from 2004 through 2006, the average up-front fee on contributions (loads) paid across investors was 24% and the average fee paid on assets under management was 26.8. These fees stood in stark contrast to the substantially lower fees on fund shares offered to independent investors, suggesting that price competition was not sufficient to lead to efficient pricing in the pension market.

We exploit policy changes in fee reporting designed to increase price sensitivity of investors to examine why investors were insensitive to management fees, how effective policy was in changing price elasticity of demand, and how the strategic response of firms impacted policy results. Halfway through our sample, the government introduced a new fee index to increase transparency of and sensitivity to management fees. The index combined fund manager load and balance fees according to a particular formula, and the government broadly advertised it to workers as the fee they should know when choosing a fund manager. We use individual-level administrative data from the Mexican social security system surrounding this policy change to test if workers insensitivity to fees stems from value placed on non-fee attributes or from a misperception of complex management fees. Because the fee index combined fees in a particular way, choosing a lower index firm could lead many workers to choose a higher-cost fund for them. We find that before the index, investors of all backgrounds paid little attention to fees when choosing fund managers. Post-policy-intervention, investors heavily weighted the fee index regardless of whether doing so caused them to choose a higher-cost fund. Investors largely ignored actual costs, choosing instead a simple-to-understand cost measure when it was made more salient by government policy.

In contrast to investors, we find that firms responded optimally to the changes in demand induced by government policy. The fee index formula over-weighted load fees and underweighted fees on assets under management. We combine our demand estimates and the formula for the fee index to show that optimizing firms should best-respond by lowering their load fees and increasing their fees on assets under management. This is in fact what they did; most firms followed this strategy which erased much of the gains to consumers from increased price sensitivity and regressively redistributed management fees from high-income to low-income segments of the market.

Overall our results add to the growing literature showing that consumers fall short of optimizing behavior, following shortcuts in lieu of complex optimization. We also demonstrate that government policy can be an effective tool for incentivizing private markets by lowering information and decision making costs to help consumers make better decisions. (Hastings and Weinstein 2008, Mastrobuoni 2009). However, we add important evidence to the literature in Behavioral Industrial Organization. We show that sophisticated firms are not fettered by the same behavioral biases that plague individual investors, and instead set fee schedules to maximize profits given investor behavioral responses to government nudges. We conclude that policies aimed at aiding consumer decision-making also need to provide the right competitive incentives for firms to be effective in increasing market efficiency.

2. Background

2.1 Overview

Mexico's privatized social security system has been in effect since July 1, 1997. The objective of the reform was to make the pension system financially viable, reduce the inequality of the previous pay-as-you-go system, and increase the coverage and amount of pensions through the establishment of individual ownership of retirement accounts. The government approved private fund administrators called Afores (Administradoras de Fondos para el Retiro) to manage the individual accounts and established CONSAR to oversee this new Sistema de Ahorro para el Retiro (System of Savings for Retirement - SAR). Six-and-half percent of wages are deposited bimonthly into the SAR account, and the worker can withdraw from this account at retirement (age 65 for men and age 60 for women), disability in old age, and for a limited amount of insurance when unemployed.[1] In June 2007, SAR had over 25 million registered accounts, and total funds in the system exceeded 1.14 trillion pesos.

[1] Mandatory contributions to the retirement account come from three places: the worker contributes a mandatory 1.125% of her base salary, the employer contributes an additional 5.15%, and the government contributes 0.225% of the base salary as well as a "social contribution" of 5.5% of the inflation-indexed Mexico City minimum wage (Sinha (2003)). Workers can withdraw unemployment insurance from the account of 1-3 months of salary depending on the amount available in the account and their contribution history. Workers must have 3 years of

During our sample period, January 2004 - December 2006, there were between twelve and twenty-one Afores in the market, with ten firms present since the inception of the system and three firms entering in the last six months of the sample. CONSAR approves each Afore's entry into the market. Afores must submit fee schedules for approval and must seek CONSAR's approval for any subsequent fee changes they wish to implement.[2]

Table 1 lists the Afores with their entry date as well as a description of the firm. The Afores range from prominent Mexican banks like Banamex to international investment firms like HSBC to department store chains like Coppel (similar to Sears); all well-known institutions in Mexico with sizable physical presence and longevity in finance, insurance or retail sectors.

2.2 Afore approval, operation and investment restrictions

During our sample, Afores were required to offer two age-based investment funds called Siefores (Specialized Investment Groups for Retirement Funds): a "higher-risk" fund for workers 55 and under called Siefore Básica 2 and a "low-risk" fund for workers over 55 called Siefore Básica 1.[3] Management fees were set at the Afore level, so the same management fee applied to both Siefores within each Afore. In addition, affiliates could not split their funds between Afores or Siefores and had to keep their funds in only one fund at one fund administrator at a time.[4]

The investment possibilities for each Siefore were heavily regulated by CONSAR. Siefore 1 was effectively restricted to investing in Mexican government bonds, and, although Siefore 2 could include investments in equities, equity investments were capped at 15% and the investment vehicles restricted to Principal Protected Notes and Exchange Traded Funds tied to major stock indices. These restrictions implied that Afores differed little on persistent performance, and tests for persistent outperformance using monthly returns show no significant difference between fund manager returns (see Appendix 1).

---

contributions to the account to qualify for unemployment insurance withdrawals. This benefit can be used one time in every five years period.

[2] Article 37, Ley de los Sistemas de Ahorro para el Retiro (Article 37, Retirement Saving System Law).

[3] In March of 2008, the system moved to a 5-fund age-based system introducing 3 'higher-risk' funds with broader investment possibilities for younger workers. See press release 08/07 for details.

[4] For these reasons we will focus our analysis on *Afore* choice since Siefore choice is completely determined by age of the worker and has no impact on relative costs.

2.3 Management Fees

Afores were allowed to charge two different types of fees, a load fee and a fee on funds under management, and despite the tight investment regulation Afores charged high and disperse management fees. The load fee was referred to as a "flow fee" because it was quoted as a percent of the worker's salary instead of as a percent of the contribution to the account, and only contributions, *not account transfers*, were subject to the load.[5] This convention implied that flow fees were reported in a way that made them seem smaller than they were - *a flow fee of 1% of salary is actually a 15.4% load fee on the contribution of 6.5% of salary (1/6.5 =0.154).* In June 2006, flow fees ranged from 0.5% - 1.65% (i.e., a 7.7% - 25.4% load). The fee on funds under management was referred as "balance fee". In addition to the flow fee, firms charged balance fees ranging from 0.12% to 1.5%.

There are two important facts to note. First, high fees are not just an artifact of social security account management costs in Mexico. Afore investments were regulated and system processes were was centralized to minimize system management costs. The management of central processes was put out for bid on multi-year contracts, and Afores paid regulated fees for centralized account processes.[6] In addition, Afores also offered shares in Seifores and account management to independent workers and for voluntary savings for retirement accounts. These identical investments had substantially lower management fees, typically keeping only the fee on assets under management and waiving the load fee (see Appendix 2).

Second, the mixture of fees between loads and balances implied that the cheapest Afore for a given worker is not necessarily the cheapest for another, because total costs depended on the wage to balance ratio of each worker. For example, a woman currently employed in the formal labor force who planned to exit the formal labor force to have children and work within the household could disregard the flow fee and choose the Afore with the lowest balance fee since she would expect to have zero contribution flow into her pension account while out of the formal labor force. The same would apply for someone exiting the formal sector to take a job in the informal sector for a sizable period of time. In Mexico, there is an active informal labor

[5] In other words, there are no monetary costs of transferring an account from one Afore to another.

[6] For example, internal information from CONSAR staff indicated that in 2008, fees for registering a new account were 25.99 Mexican pesos, 0.62 pesos for processing each contribution into the account, and 5.47 pesos for each switching of accounts (fee charged to the Afore accepting the account). One dollar is approximately 12 Mexican pesos.

sector with 30% of workers with college education (overall 27% of investors) spending time in both the formal and informal employment sectors from 2005 through 2010, and 60% of workers with non-college backgrounds spending time in both sectors over the same time period.[7] Forward-looking agents with full information should take advantage of relative fee changes and move to a fund manager with zero balance fees upon leaving the formal sector. This provides variation in the relative management costs of each Afore as well as a simple test for forward-looking behavior.

In addition, further variation in management costs was induced by a change in regulatory policy towards tenure discounts. Most firms offered a small tenure discount off of the flow fee for clients that had passed a certain tenure point with the Afore. This was typically a basis point discount per year of tenure, making other Afores relatively more expensive as tenure increased. In January 2005, CONSAR required that all tenure discounts be recalculated based on years in the system instead of years with an Afore, erasing the built-in switching costs.[8] This regulatory change caused further differential shocks to relative fund manager expenses across individuals during the first half of our sample period.

3. Regulatory Changes, Information and Management Fees

Of course, multiple fees, discounts, and changes in discount rules make it more difficult to calculate the alternative costs of each Afore. To simplify fee information for affiliates, CONSAR created a composite fee index called the "Equivalent Fee on the Balance", and beginning in July of 2005 made a push to publicize this fee as the fee workers should be looking for when choosing an Afore. We will refer to this fee as the CEF (CONSAR's Equivalent Fee). The index was constructed in the following way: calculate the accrued balance for a person with wage W, balance B, and tenure T at the end of time horizon H at each Afore's current flow and balance fees and a real rate of return (assumed uniform across Afores at 5%), then calculate the balance fee that would lead to the same balance if flow fees were set to zero. This is the Equivalent Fee on the Balance, and it is expressed as an annual percentage rate.

[7] Based on author's calculations from the 2010-2011 Encuesta de Empleo Retiro y Ahorro, a survey of SAR account holders in Distrito Federal.
[8] See CONSAR press release 07/05 for details.

Prior to July 2005, CONSAR calculated this fee using a 25 year horizon which implied that the 25 year CEF (CEF25) was close in magnitude to balance fees. Differences between Afores in the CEF25 also appeared small in absolute value even though these small differences imply large differences in account balance when compounded over 25 years.

From July 2005 onward, CONSAR mandated that the CEF be computed over a 1 year period (CEF1) instead of over a 25 year period. This tripled the size of the CEF, making it closer in size to a flow fee (as a percent of wage) than the balance fee, and increased the absolute fee difference between the Afores. In addition to changing the CEF used from the CEF25 to the CEF1, CONSAR also introduced new regulations requiring the prominent display of a comparative CEF1 table on the front page of each worker's account statement.[9] Moreover, they also required that each affiliate sign a form stating that (he or she) saw and understood the CEF1 table when submitting an application to switch Afores, potentially harnessing Afore sales force to advertise the CEF1 when recruiting customers.

Table 2, columns 1 through 3 show the flow fee as quoted (a percent of salary), the implied load as a percent of contributions (flow fee / 6.5), and the balance fee for each Afore in June 2005, on the eve of the CEF1 introduction and information mandate. The table is sorted in ascending order by CEF25 (column 6), with Actinver at the top with a CEF25 of 0.55 and Profuturo last with a CEF of 1.14. Columns 4 and 5 show the share of account holders and assets under management in each Afore as of June 2005. Note that larger share firms are located in the lower half of the table, and firms like Santander and Banamex are dominated on both fee dimensions by other firms, yet have larger market shares.

Columns 6, 7 and 8 show the CEF25, the CEF1 and the rank of the Afore according to the CEF1. Note the size of the CEF increases 3 to 5 fold when the one year amortization is used. In addition, the relative ranking of the firms based on the CEF changes substantially, even though the underlying management fees used for the calculation are unchanged. Thus changing the CEF formula could have resulted in a large change in perceived management fees even though the actual fees were unchanged.

[9] See press release 10/05 for details.

4. Empirical Analysis of Response to Information

4.1 Descriptive Statistics on Consumers

We construct a panel data set for investors and firms from raw administrative data from January 2004 through December of 2006 which records labor force participation, earnings, mandatory contributions to retirement accounts, account balances and switches between Afores for all account holders from the inception of the system through the end of 2006. We combine this with a monthly panel of Afore fees and a constructed history of regulatory changes published in official government registries to measure the impact the information intervention had on investment choices, demand for fund managers, and firm pricing strategies.[10]

We begin by looking at raw data on movement of accounts between Afores before and after the policy intervention. Table 2 showed that Afores with large market shares were dominated by Afores with lower market shares on both price dimensions. This could be because individuals actively choose higher priced Afores because they place high value on non-price attributes, because individuals choose low-cost Afores at the time of choice but update infrequently relative to price changes, or because individuals cannot easily measure management fees or their relationship to wealth at retirement.

Table 3 summarizes movements in accounts between Afores at the time of choice before and after the policy intervention. It shows how investor decisions changed before and after the introduction of the CEF1 as the official fee index with respect to movements along alternative measures of price and management cost. We construct several measures of management costs and examine whether investors were moving from higher-cost to lower-cost Afores along each of these measures before (columns 1 through 3) versus after the information intervention (columns 4 through 6). In the first five rows we compare investor movements between Afores based on the Afore's relative CEF, flow and balance fees. For example, if we rank each Afore based on the official CEF (CEF25 pre-intervention, and CEF1 post-intervention), with lower ranks representing lower CEF's, we see that pre-intervention, the median investor was moving to an Afore with a 2-rank-higher CEF25 than the one they were currently with. Twenty-five percent of

[10] These include press releases from CONSAR available HERE, Circulares available HERE, and the original Ley del SAR from 1997.

investors were moving to Afores with 3 rank lower CEF25's, however twenty-five percent were moving to Afores with implying a 5-rank increase in the CEF25.

After the information intervention, these statistics change sharply, with the median investor moving to a 2-rank-lower Afore than their current Afore as measured by the new CEF1. The change occurs at all points in the distribution, with 25% of investors moving to an Afore with a size-rank lower CEF1, and the 75th percentile measure decline as well from 5-ranks higher to 3-ranks higher.

Because the CEFs are functions of flow and balance fees, we can see similar movements in flow versus balance fees. Comparing rows 2 and 4, we see that the median investor was moving to an Afore with 0.01 higher flow fee than their current Afore. Had they moved to the Afore with the cheapest flow fee, they could have saved 1.10 percentage points off the flow fee, or 16.9% (1.10/6.5) of their salary contributions. Even the quartile with the least 'flow fee' left on the table could still have saved 13.8% (0.90/6.5) of their salary by switching to the lowest flow fee Afore. Similarly, rows 3 and 5 show that the median worker was moving to an Afore with a slightly higher balance fee, and leaving 40 basis points in annual fees on assets under management fees on the table. Post information intervention, these statistics improve: the movement towards lower-CEF1 Afores was associated with movements to Afores with slightly lower flow and balance fees than before the information-intervention.

However, as described earlier, many workers may move to a high flow fee Afore with zero balance fee if they plan to work in the informal sector or exit the work force for a while. What is the cheapest Afore for one worker may not be the cheapest Afore for another, thus looking independently at aggregate statistics on each fee may mask sensitivity to actual management costs, and the change in sensitivity to the CEF1 could be generated by the fact that the CEF1 is more closely aligned than the CEF25.

To examine changes in management costs, we construct three cost measures. One is a present discounted value of cost until retirement based on each individual's average wage and formal-sector employment rate over our three year period. The second is a predicted cost measure which uses actual baseline formal-sector employment and wages at the time of switching to construct an expected wage and formal-sector employment rate going forward based on individuals with very similar baseline characteristics (age, system tenure, gender, historic employment rate, balance and wage). This is like a regression prediction of management costs

(see for example Abaluck and Gruber (2011)). The third is a myopic cost measure which assumes that the individual's current employment status and wage at the time of switching is what they expect going forward.

For each of these cost measures, we convert the present discounted value of management costs into days of current wages to facilitate comparison across individuals. Rows 6 through 8 show the change in management costs in days between the old and new Afore before and after the information intervention. Rows 9 through 11 show how many days of wages could have been saved if the individual switched to the lowest-cost Afore for them rather than the Afore they chose. Prior to the information intervention, the median person was switching to an Afore that cost them between 7 and 16 days of wages more in management fees than their current Afore. Post information intervention, this changed with the median person switching to an Afore with 3.8 to 8.8 lower days-cost than their current Afore. In addition, more workers were choosing lower-cost Afores post-intervention than before the intervention, though a substantial fraction were still actively moving towards higher-cost Afores according to each of the cost measures. In fact, rows 9-11 show that almost all workers were continuing to leave a substantial amount of money on the table. The median worker could have saved close to 120 days of wages by choosing the cheapest Afore for them pre-intervention, and close to 100 days of wages post-intervention.

Table 3 shows that workers sought lower CEF Afores post-intervention, but this led to only modest declines in management cost savings for most workers. This can be explained in part by the fact that the CEF1 was not reflective of actual management costs, as it only considered 1 year costs which placed a very high weight on the flow fee relative to fees on assets under management. In addition, it assumed a particular annual formal-sector contributions, account balance, and tenure in the system which was not reflective of most individuals in the system. Thus it could direct many individuals towards higher cost Afores. To illustrate this point, Figure 1 shows the share of account movements pre- and post- intervention that moved to each combination of higher/lower cost/CEF Afores. Pre-intervention, 42.5% of switchers moved to a lower-CEF25 Afore, but post-intervention, this number jumped to 63.5%. However of that 63.5%, over a third of them (23.1%) moved to an Afore with higher expected costs for them. This is due to the fact that one year costs shift individuals to Afores with low flow fees even if those flow fees are irrelevant to them. On average, though, because most people expect positive

account flows, the increased focus on the CEF caused by the information intervention resulted in more people overall moving to lower-cost-for-them Afores (55.8% versus 44.1%)

The fact that investors responded to the information intervention by seeking lower CEF1 Afores even if that led them to choose higher-expected-cost Afores suggests that the response to the CEF1 was caused by investors following shortcuts as substitutes for costly optimization. If price insensitivity apparent in summary statistics on account movements was caused by preferences for non-price attributes then government price indices should have no impact on overall choice behavior. If investors correctly understand the index, they should ignore it if it is inversely correlated with their own expected management costs.

To further examine choice behavior and the impact of the CEF, Tables 4 and 5 split the sample by formal-sector employment. Table 4 shows summary statistics on switching before and after the CEF1 for those who were always employed in the formal sector versus those who were never employed in the formal sector during our sample. First, for those who are never employed in our sample, we would expect them to move towards lower balance fee funds since they are unlikely to make contributions for at least a few years. Looking at the pre-intervention period, and comparing those who are always employed versus those who are never employed in the formal sector, we find little evidence that minimizing management fees is driven by workers' labor market participation. In addition, post-intervention, both types of workers change their behavior and choose Afores with lower CEF1's, despite the fact that a lower-CEF1 Afore is more likely to have a lower flow fee than a lower balance fee making them potentially higher cost for those not formally employed. This is reflected in the fact that post-intervention CEF1 focus resulted in a significant expected savings in management costs for those always formally employed, but not for those who were never formally employed.

Table 5 repeats Table 4 but focuses only on individuals with a last-recorded-formal sector wage in the top quartile of wage earners. We use this as a proxy for education of the worker to examine if those who are likely highly-educated and always unemployed choose Afores to minimize personal management costs. Again, even among this group, we find little difference in behavior between those always and never formally employed during our 3 year sample. If anything, those never employed appear to move towards lower flow-fee Afores pre-intervention (the wrong fee to choose on), and both types of workers appear to choose lower CEF1 Afores

post-intervention even though that leads to no average cost savings gains for those in the informal sector.

Table 6 splits the sample by age and by length of participation in the system (time since first formal sector wage contribution under the new 1997 privatized social security system). First, workers of all ages shift from choosing higher-CEF Afores to lower-CEF Afores post intervention. Hence the information intervention appears to have shifted the choice behavior of participants of all ages. Second, young workers seem to be the worst decision makers, actively choosing higher fee funds pre-intervention. Because young workers have low balance to wage ratios by definition, shifting them towards a CEF1 which weights flow fees more than balance fees could impact them positively. Indeed, the intervention moved them towards lower expected cost fund managers, while it was fairly neutral for older workers who may move to higher or lower cost fund managers if they select based on the CEF1 since it is unclear if they should weight a flow fee more than a balance fee. Similar results hold based on experience in the system. Those with lower experience by definition have high wages relative to balances, but sought higher-flow fee funds managers before the information intervention. These workers moved from being cost-loving (if anything) to at least cost-neutral as a result of the information intervention.

Thus it appears that younger workers and inexperienced workers made relatively worse decisions pre-intervention, and gained the most from seeking CEF1 post-intervention as the CEF1 was more closely aligned with their wage to balance profile. Looking across subgroups, it is hard to find a subgroup of workers that chose Afores to minimize personal costs based on complex fee structures and personal wage to balance expectations pre-intervention. All subgroups appear to have followed the new CEF1 post-intervention as a short-cut or cost approximation regardless of whether it led them to higher- or lower-cost-for-them Afores.

4.1 Descriptive Statistics on Firms

The government information “nudge” appeared to be effective at shifting demand, but towards a measure that was not necessarily positively correlated with management costs for many workers in the system. Firms were effectively required to advertise this fee index as it was mandated to be displayed on the front page of each statement, and in every switching transaction

from one Afore to another. Thus the information intervention may have been successful in part because it harnessed the sales force of Afores to advertise it.

In fact, profit maximizing Afores may not have protested the fee index if it allowed them to rebalance their fee structure to increase profits while attracting customers who were seeking lower fee index funds. In particular, if the information intervention made workers sharply more sensitive to the CEF, but not necessarily to actual management costs, we would expect to see a change in firm pricing if firms correctly optimize against demand and how the CEF is affected by flow and balance fees. Firm response gives us an additional supply side moment identifying the change in demand.

To see how the information intervention and resulting change in demand would impact firm's incentives, we can write the profit for Afore *j* at time *t* as

$$\Pi_{jt} = \sum_{i=1}^{N_t} \boldsymbol{rev_{ijt}}(\boldsymbol{f_{ijt}}, \boldsymbol{b_{jt}}, \boldsymbol{X_{it}}) - \boldsymbol{mc_{it}} * \boldsymbol{q_{ijt}}(\boldsymbol{CEF_{jt}}(\boldsymbol{f_{jt}}, \boldsymbol{b_{jt}}, \boldsymbol{s_t}), \boldsymbol{CEF_{-jt}}, \boldsymbol{cost_{ijt}}(\boldsymbol{f_{ijt}}, \boldsymbol{b_{jt}}, \boldsymbol{X_{it}}), \boldsymbol{cost_{i-jt}}, \boldsymbol{\delta_t}; \boldsymbol{\theta_i}) \quad (1)$$

where rev_{ijt} is the revenue for Afore *j* from person *i* at time *t*, which is a function of the flow fee firm *j* charges person *i* at time *t* (inclusive of tenure discounts and tenure discount regulatory changes), the balance fee *j* charges at time *t*, and the contribution and balance characteristics of person *i* at time *t*, X_{it}; mc_{it} is the marginal cost of managing *i's* account at time *t*; q_{ijt} is the probability that person *i* selects *j* (either actively or passively) to manage his or her account at time *t*. The choice of Afore is a function of the CEF, which itself is a function of flow and balance fees, and assumptions (supuestos) placed on CEF calculation, s_t, the expected management cost for *i* in each Afore *j* at time *t*, non-price characteristics of the Afores, $\boldsymbol{\delta_t}$, and individual specific preferences for Afore characteristics, θ_i.

The probability of choosing an Afore, $q_{ijt,}$ is a function of the CEF as well as management costs, which are a function of flow fees, balance fees and worker's characteristics. It is also a function of preferences, or the relative weight workers place on management fees and the CEF. The change in demand with respect to the two fees Afores charge is:

$$\frac{\partial q_{ij}}{\partial f_j} = \frac{\partial q_{ij}}{\partial CEF_j} * \frac{\partial CEF_j}{\partial f_j} + \frac{\partial q_{ij}}{\partial cost_{ij}} * \frac{\partial cost_{ij}}{\partial f_j}$$

$$\frac{\partial q_{ij}}{\partial b_j} = \frac{\partial q_{ij}}{\partial CEF_j} * \frac{\partial CEF_j}{\partial b_j} + \frac{\partial q_{ij}}{\partial cost_{ij}} * \frac{\partial cost_{ij}}{\partial b_j}$$

Table 7 shows how $\frac{\partial CEF_j}{\partial f_j}$ and $\frac{\partial CEF_j}{\partial b_j}$ changed with the change from the CEF25 to the CEF1. The derivatives are evaluated at the fees in place right before the information intervention intervention; in June of 2005. Note that the responsiveness of the CEF to the flow fee quadruples, while the responsiveness to the balance fee decreases slightly. In particular, this means that an Afore could lower their flow fee by one percentage point, and raise their balance fee by four percentage points and their CEF1 would remain the same. Under the CEF25, this same restructuring would have resulted in a substantially higher CEF. Furthermore, under the new CEF1, a firm could have lowered their flow fee by one percentage point and raised their balance fee by three percentage points to achieve a lower CEF1. Depending on the characteristics of their existing clients and the clients they would most likely attract with a lower CEF1, this reduction in CEF1 could have resulted in higher revenues per customer while at the same time attracting customers who were seeking a lower fee index.

Tables 3-7 show in the raw data that investors of all backgrounds moved to lower CEF1 Afores post information change, and that the CEF1 could be lowered by simultaneously lowering the flow fee and raising the balance fee, mitigating any revenue losses from the lower flow fee, and perhaps even resulting in higher revenues. If investors became much more sensitive to the CEF and the CEF became much more sensitive to the flow fee, Afores could find it profitable to lower flow fees and raise balance fees following the information intervention. Table 8 shows that this is indeed what occurred. Prior to June 2005, most Afores followed a high-flow-low-balance strategy with fees remaining relatively flat over time. Following the information intervention, fees changed dramatically. By the end of 2007, most Afores had dropped their flow fees and raised their balance fees by multiples with the effect of considerably lowering their CEF1.

To formally link firm response to information intervention and demand response, we estimate demand for fund managers as a function of the CEF and management costs from January 2004 through December of 2006. We then use data on all account holders to calculate

each Afore's best response flow fee and balance fee to the information intervention and resulting change in demand given characteristics of their current account holders.

4.3 Model of Demand

We estimate a random utility model of demand for Afores where workers choose Afores, *j*, to maximize utility function as a function of expected management costs, $cost_{ijt}$, the CEF_{jt}, and Afore-specific values, $v_{ijt.}$

$$\boldsymbol{u_{ijt} = \alpha_{ijt} cost_{ijt} + \gamma_{ijt} CEF_{jt} + v_{ijt}} \quad (2)$$

To tractably allow for preference heterogeneity, we estimate this model separately pre- and post-information intervention, setting the CEF equal to the CEF25 pre-intervention and the CEF1 post-intervention. Within each regulatory period, we estimate a conditional logit model separately by age quartile, wage quartile, and gender, allowing preferences for all Afore characteristics to fully interact with these demographic characteristics.

$$\boldsymbol{u_{cjt} = \alpha_{cjt} cost_{ijt} + \gamma_{cjt} CEF_{jt} + \delta_{cjt} + \varepsilon_{ijt}} \quad (3)$$

Where *c* indexes the demographic and regional cell that individual *i* falls into, *j* indexes the Afore, *t* indexes the pre-intervention versus post-intervention periods, δ_{cjt} is a cell-time period mean valuation for afore *j* which captures mean observable or unobservable characteristics of the Afore such as expected future returns, prevalence of branches, friendliness of service, etc., and ε_{ijt} is an i.i.d. extreme value error term.

Coefficients on management costs versus coefficients on the CEF are identified in several ways. First, changes in "supuestos" - assumptions placed on the balance, wage, tenure and minimum wage level used in the CEF formulas - cause periodic changes in the CEF's of the Afores independently from changes in underlying fees. Second, there are a handful of entries and exits during the pre-period changing the choice set. Third, regulations in the pre-period changed how discounts for tenure were applied, exogenously changing the relative flow fees of each

Afore based on system versus Afore tenure of each individual and the Afore's predetermined discount policy. Fourth, changes in fees change the costs versus the CEF in different ways for different workers based on how different the worker is from the "supuestos" used to make the CEF. Finally, even conditional on demographics and area of residence, workers will face different costs at each Afore based on their incoming estimated endowment balance when the system privatized, and based on their relative expected time spend inside or outside of the formal sector. In each specification, estimated sensitivities to CEF versus cost are robust to the inclusion of one or both of the fee measures in the utility function, implying that the impact of each on demand is separately identified.

Table 9 shows summary statistics for demand elasiticities with respect to cost versus the CEF pre- and post-intervention. We evaluate the elasticities at the estimated parameter on the full estimation sample. Pre-intervention, the average elasticity with respect to the CEF across all individuals was negative but clearly less than one in absolute value. Post information intervention, the average is close to if not over one for every Afore, with the exception of Inbursa, the financial arm of Telemex owned by Carlos Slim, which had not changed its fee structure (.50, .50) for a decade and did not change fees in response the CEF change.[11] Elasticities with respect to expected management costs were near zero before and after the information intervention.

These elasticity estimates echo the changes in mean flow of accounts from Table 3; suggesting that investors became more sensitive to the CEF post-intervention, and that Afores lowered the CEF1 in response to this increased investor sensitivity to compete for individuals who were considering switching Afores. Because investor elasticity with respect to management costs remained near zero, Afores could lower their CEF1 by raising balance fees and lowering flow fees (as opposed to lowering both fees) without adversely impacting demand even if this resulted in higher management costs.

[11] Because our model controls for Afore fixed effects by demographic group pre and post intervention, endogeneity of prices would need to occur because of changes in the value of unobservable Afore characteristics over time changed within time period in manner correlated with Afore changes in flow, balance and therefore CEF. It seems unlikely that this would occur and bias our estimates of changes in price sensitivity post information intervention. Nonetheless we estimate the post intervention demand parameters instrumenting for price (and estimated cost) using a calculated best response flow and balance fee for each Afore assuming preferences remain unchanged but the CEF formula changed to the CEF1. This uses only baseline characteristics of investors and Afores in the pre-intervention period to estimate demand parameters post-intervention. We find qualitatively similar results. See section 4.4, equation 4, footnote 14 and Appendix 4 for details on the calculation of best responses.

4.4 Model of Supply

Given the demand estimates, we can calculate Afore best-responses to the policy intervention and the change in consumer preferences to formally link the fee restructuring to the policy change. Precisely we estimate the expected present discounted value of profits given the market, current account holder characteristics, and demand as of June 2005, the eve of the policy reform. We evaluate the profit maximizing flow and balance fee subject to the constraint that any fee combination would have to result in a *lower* official CEF (CEF25 pre-intervention and CEF1 post-intervention) than the CEF of their current fee schedule.[12]

The present discounted value of profits for each Afore can be written as (see more detail in Appendix F),

$$\boldsymbol{\Pi_j = \sum_t \left[\delta^t M_{it} * rev_{i,jt} * S_{0,i,j} \prod_{k=0}^{t} (1 - \alpha_{ik}) + M_{it} * rev_{i,jt} \delta^t \sum_{l=0}^{t-1} \{ \alpha_{il} q_{ijl} \prod_{k=0}^{n-l-1} (1 - \alpha_{in-k}) \} + \delta^t M_{it} \alpha_{it} * q_{jit} * rev_{i,jt} \right]} \quad (4)$$

where , α_{it}, is the probability that an individual *i* evaluates her savings and retirement account and her Afore choice in any time period *t*, $S_{0,i,j}$ is an indicator if person i is an affiliate of Afore *j* in June 2005 (time 0), rev_{ijt} is the expected revenue that Afore *j* will receive from i given *j*'s fee structure and *i*'s characteristics and preferences at time *t*,q_{jit}, is the demand for Afore *j* at time *t* and δ is a discount rate.

To simplify the analysis we will assume: i) the probability α_{it}, is constant over time for all types of individuals, *i*, ii) preferences governing demand for Afores are also the same over time within individual of type *i*, iii) revenues are constant over time and iv) retirement out of and entry into the savings and retirement market is constant over time. Then the final equation for the discounted value of profits for each Afore is,

[12] Given the inelastic demand in the Pre-period, Afores could have charged higher prices in equilibrium in the absence of regulatory review and implicit fee regulation. We model this regulatory pressure as implying any new fee schedule would need to result in a lower official fee index or it would not be approved by the regulator.

$$\Pi_{ij} = M_i rev_{ij} S_{ij0}(1 - \alpha_i) \sum_{t=0}^{\mathrm{T}} \left(\delta(1 - \alpha_i)\right)^t + M_i rev_{ij} \alpha_i q_{ij} \sum_{t=0}^{\mathrm{T}} \delta^t \sum_{l=1}^{t} (1 - \alpha_i)^l + M_i rev_{ij} \alpha_i q_{ij} \sum_{t=0}^{\mathrm{T}} \delta^t \quad (5)$$

Where, α_i is the probability that an individual *i* evaluates her savings and retirement account and her Afore choice in any period, $S_{0,i,j}$ is an indicator if person *i* is an affiliate of Afore *j* in June 2005 (time 0), rev_{ijt} is the expected revenue that Afore *j* will receive from *i* given *j's* fee and q_{ij} is the demand for Afore *j*. Appendix F derives this profit function in detail.

The first term of the profit function is the net present discounted value of revenue the Afore receives from its current (June 2005) client base who never evaluate their account. These individuals will choose Afore *j* no matter what the fee is as they never 'wake up' to evaluate their account. The average worker in fact has never switched Afores from the original Afore they signed up with, and only 10 percent of account holders switch per year. Appendix 3 presents estimates from a discrete time hazard model of Afore switching and demonstrates that the single largest determinant of Afore switching is employment status: active workers in the formal sector are more likely to evaluate their accounts and periodically switch Afores, while workers who are inactive and no-longer making contributions to their account for more than 6 months are very unlikely to switch fund managers. Thus the inframarginal, or 'captive' account holders are unlikely to have flow fee revenues, but likely to have balance fee revenues. Thus lowering flow fees but raising balance fees could both increase revenues on inframarginal clients and attract new clients by lowering the CEF1.

The second term is the revenue from individuals who evaluate their account at date, *t* , and choose afore *j* with probability q_{ijt}according to *j's* characteristics and their preferences over those characteristics at time *t*., but then do not ever evaluate their account again until a later future date, thus adding to *j*'s inframarginal consumer base in future dates. The third term is revenues gained each period from those who evaluate their Afore choice and decide to select afore *j* with probability q_{ij}.

Thus an Afore's profit is affected by fees through the impact on revenues for current clients who are not paying attention to their accounts and through the impact on expected revenues and expected demand response for those evaluating their accounts in a given period and switching to Afore *j* based on their preferences for its relative fees and non-fee characteristics.

We calculate profits for each Afore at alternative fee structures holding the other Afore's prices fixed to examine how the demand change and the CEF formula change affects their best response function. Further, we assume that Afores optimize assuming that current consumer preferences and account evaluation behavior will remain the same in the future.

We calculate this profit function for each Afore on a 0.10 grid of balance and flow fees evaluated at the CEF25 formula and the pre-intervention demand estimates, and then the CEF1 formula and the post-intervention demand estimates. We calculate a grid rather than an analytic first order condition as the profit function may not be differentiable on the set of possible fees due to the inelastic base of inframarginal customers (Hastings, Hortacsu and Syverson 2011). Appendix 4 outlines our calculation approach in detail.

Table 10 shows these calculations for each Afore. In the pre-intervention period we find that if anything, Afores should lower balance fees and raise flow fees from their current levels if they were charging any balance fee at all.[13] However, the calculations imply that Afores should have high flow fees and low balance fees given the pre-period CEF25, estimated preferences, and investor characteristics. After the information intervention, this switches dramatically. Afores now have the incentive to drop flow fees to zero and increase balance fees several fold, in line with the behavior that we see.[14]

Higher balance fees and lower flow fees would benefit workers with low balances relative to inflows. Table 11 calculates the redistributive and overall impacts of the policy on management costs. To do this we compare expected revenues for each Afore at their June 2005 fees and their December 2007 fees using the same formula we used to calculate the best responses to the policy change, under the assumption that fees by December 2007 are at a new equilibrium. Table 11 shows that the move to the CEF1 with the accompanying response by consumers and firms resulted in an overall reduction in management costs but a redistribution of costs from wealthier to lower-income affiliates. This is largely due to the fact that low income affiliates are more likely to spend time out of the formal sector, and are less likely to periodically evaluate their accounts and switch Afores to minimize management fees. They are an inelastic

[13] We might get this deviation from actual fees as the profit function is approximate and evaluated using universal administrative data that Afores do not have access to. It is an open question as to how firms optimize when demand is not fully known.

[14] In fact the change in the CEF formula alone turns out to be sufficient to generate this response. If we do the same profit calculations in the post-intervention period using the new CEF1 formula but holding preferences constant at their pre-intervention levels, we find the same change in incentives for Afores. Their best responses indicate setting flow fees to zero and substantially increasing balance fees.

group for whom management costs form balances outweigh management costs from fees on wage contributions. Had the index increased elasticity for marginal customers (high wage earners) without distorting the relative importance of load versus balance fees and thus firm strategy, this redistribution would have been smaller.

5. Conclusion

We use a unique and detailed data set on administrative records in Mexico's privatized social security system surrounding a major information-intervention to test if workers understand management fees and act to maximize utility with wealth at retirement a primary determinant of choice. We show that investors are not sensitive to actual management costs, but instead choose fund managers based on short cuts such as summary indices of fees published by the government. They focus on fee indices even if this leads them to choose fund managers with higher expected costs for them. We find that workers from all backgrounds change their choice behavior dramatically when the government introduces a new fee index and advertises it as a measure of cost. Worker's who's characteristics match those used for the fee index are helped by the index, while those whose characteristics do not match the index choose fund managers based on the index even if it is irrelevant for them and even if it leads them to choose a higher-cost fund manager.

In contrast to fettered investors, firms are sophisticated. We show that firms had an incentive to reweight their fees in response to the government's information intervention. Their fees change accordingly, allowing them to capture consumers who seek a lower index, but increase revenues per customer simultaneously by raising less-salient fees when they lower the fees most salient in the government's fee index measure.

This paper adds to growing evidence that consumers may not be the best agents for themselves when faced with complex decisions with delayed payoffs. They are likely to respond to short-cuts such as salient fees, suggestions or advertising. The government can assist the functioning of markets by creating short cuts that facilitate consumer choice, but those short cuts cannot be easily gamed by sophisticated firms who do not face such behavioral biases.

References

Abaluck, Jason and Jonathan Gruber. (2009). "Choice Inconsistencies Among the Elderly: Evidence from Plan Choice in the Medicare Part D Program." *NBER Working Papers 14759.*

Ausubel, Lawrence M. (1991). "The Failure of Competition in the Credit Card Market," *American Economic Review*, 81(1), 50-81.

Ashraf, Nava, Dean Karlan and Wesley Yin. (2006). "Tying Odysseus tot he Mast; Evidence from a Commitment Savings Product in the Philippines." *The Quarterly Journal of Economics*. 121(2), 635-672.

Bagwell, Kyle. (2007). "The Economic Analysis of Advertising," Handbook of Industrial Organization. Elsevier.

Barber, Brad, Terrence Odean and Lu Zheng. (2005). "Out of Sight, Out of Mind: The Effects of Expenses on Mutual Fund Flows." *Journal of Business*, 78, 2095-2119.

Bernartzi, Shlomo and Richard Thaler. (2004). "Save More Tomorrow: Using Behavioral Economics to Increase Employee Savings." *Journal of Political Economy*,

Bernartzi, Shlomo and Richard Thaler. (2001). "Naive Diversification Strategies in Retirement Saving Plans." *American Economic Review*, 91(1), 79-98.

Bollinger, Bryan, Leslie. P. and A. Sorensen. (2009). "Calorie Posting in Chain Restaurants". Unpublished Working Paper.

Beshears, John, James J. Choi, David Laibson and Brigitte C. Madrian. (2006). "Simplification and Saving." *NBER Working Papers 12659.*

Carroll, Gabriel, James Choi, David Laibson, Brigitte C. Madrian, and Andrew Metrick. (2009). "Optimal Defaults and Active Decisions." *The Quarterly Journal of Economics*, forthcoming.

Chetty, Raj, Adam Looney and Kory Kroft. (2007). "Salience and Taxation: Theroy and Evidence." *NBER Working Paper 13330.*

Chevalier, Judith and Glenn Ellison. (1997). "Risk Taking by Mutual Funds as a Response to Incentives." *Journal of Political Economy*, 105, 1167-120.

Choi, James J., David Laibson and Brigitte C. Madrian. (2007). "$100 Bills on the Sidewalk: Suboptimal Investment in 401(k) Plans." *NBER Working Paper 11554.*

Choi, James J., David Laibson and Brigitte C. Madrian. (2006). "Why Does the Law of One Price Fail? An Experiment on Index Mutual Funds." *NBER Working Papers 12261*.

Cronqvist, Henrik. (2006). "Advertising and Portfolio Choice." *Unpublished Working Paper.*

Cronqvist, Henrik and Richard Thaler (2004). "Design choices in privatized social-security systems: Learning from the Swedish experience." *American Economic Review* 94, 424-428.

DellaVigna, Stefano. (2009). "Psychology and Economics: Evidence from The Field." *Journal of Economic Literature*, 47, 315-372.

Dixit, A. and V. Norman. (1978). "Advertising and Welfare." *The Bell Journal of Economics*, 9, 1-17.

Dorfman, R. and P. O. Steiner. (1954). "Optimal Advertising and Optimal Quality." *American Economic Review*, 44, 826-36.

Duflo, Esther and Emmanuel Saez. (2003). "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence From a Randomized Experiment." *The Quarterly Journal of Economics*, 118, 815-842.

Ellison, Glenn. (2006). "Bounded Rationality in Industrial Organization." (eds) Richard Blundell, Whitney Newey, and Torsten Persson. Advances in Economics and Econometrics: Theory and Applications. Cambridge University Press.

Farrell, Joseph and Paul Klemperer. (2007). "Coordination and Lock-In: Competition with Switching Costs and Network Effects." Handbook of Industrial Organization. Elsevier.

Gabaix, Xavier and David Laibson. (2006). "Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets." *The Quarterly Journal of Economics*, 121, 505-540.

Hastings, Justine S., and Lydia Tejeda-Ashton. (2008). "Financial Literacy, Information, and Demand Elasticity: Survey and Experimental Evidence." *NBER Working Paper No. 14538*.

Hastings, Justine S., and Jeffrey M. Weinstein. (2008). "Information, School Choice and Academic Achievement: Evidence from Two Experiments," with Jeffrey M. Weinstein. Quarterly Journal of Economics, 123(4), 915-937.

Hortaçsu, Ali and Chad Syverson. (2004). "Product Differentiation, Search Costs, And Competition in the Mutual Fund Industry: A Case Study of S&P 500 Index Funds." *The Quarterly Journal of Economics*, 119, 403-456.

Katz, Michael L and Carl Shapiro. (1994). "Systems Competition and Network Effects." *Journal of Economic Perspectives*, 8, 93-115.

Klemperer, Paul. (1995). "Competition when Consumers have Switching Costs." *Review of Economic Studies*, 62, 515-39.

Kling, Jeffrey R., Sendhil Mullainathan, Eldar Shafir, Lee Vermeulen, and Marian V. Wrobel. (2008). "Confusion and Choice in Medicare Drug Plan Selection." Unpublished Manuscript.

Madrian, Brigitte and Dennis F. Shea. (2001). "The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior." *The Quarterly Journal of Economics*, 116, 1149-1187.

McFadden, Daniel. (2006). "Free Markets and Fettered Consumers (Presidential Address to the American Economic Association)." *American Economic Review*, 96(1), 5-29.Sirri, Erik and Peter Tufano. (1998). "Costly Search and Mutual Fund Flows." *Journal of Finance*, 53, 1589-1622.

Thaler, Richard and Cass Sunstein. (2008). Nudge: Improving Decisions about Health, Wealth and Happiness. Yale University Press.

TABLE 1: ENTRY DATES AND DESCRIPTIONS OF AFORES PRESENT IN MARKET DURING SAMPLE PERIOD, JANUARY 2004 – DECEMBER 2006

Afore Name	Entry Date	Firm Description and Brand Perception
Actinver	Apr-03	Mexican financial group
Afirme Bajío	Dec-05	Mexican financial group
Ahorra Ahora	Aug-06	Owned by Mexican financial group Monex
Argos	Dec-06	Mexican insurance company affiliated with international insurance company Aegon
Azteca	Mar-03	Grupo Salinas (owns Elektra retailer for low- to middle-income demographic groups and the TV chain Azteca)
Banamex	Jul-97	Large Mexican bank (since 1884), bought by Citigroup (2001)
Bancomer	Jul-97	Large Mexican bank (since 1932), affiliated to Spanish Bank (in 2000)
Banorte Generali	Jul-97	Northern Mexican bank affiliated with International Insurance Company Generali
Coppel	Apr-06	Mexican leading departmental store for low- to middle-income demographic groups
De la Gente	Nov-06	Joint venture of small savings institutions and government bank (BANSEFI)
HSBC*	Jul-97	International Bank
Inbursa	Jul-97	Banking and financial services group (owned by Carlos Slim)
ING**	Jul-97	International financial group
Invercap	Feb-05	Mexican mutual funds administrator founded in the north of Mexico
IXE	Jun-04	Mexican financial group
Metlife	Feb-05	International insurance company
Principal	Jul-97	International financial group
Profuturo GNP	Jul-97	Mexican mutual funds administrator
Santander	Jul-97	Spanish bank that bought the Mexican Bank Serfin in 2000
Scotia	Nov-06	International banking and financial services company
XXI	Jul-97	Owned by IMSS (former pension system administrator) and Prudential

*HSBC acquired Afore Alianz Dresdner in 2004 which was Afore Bancrecer Dresdner until 2001.**ING acquired Afore Bital in 2001. Bital is a Mexican bank.

TABLE 2: AFORE FEES AND MARKET SHARE BY FEE INDEX PRE- AND POST-INTERVENTION

Afore Name	Flow Fee	Balance Fee	Share Accounts	Share Assets	25 Year CEF	1 Year CEF	Rank 1 Year CEF
Actinver	1.03	0.20	0.001	0.002	0.55	2.02	2
Azteca	1.10	0.15	0.003	0.005	0.58	2.22	4
Invercap	1.03	0.20	0.000	0.000	0.60	2.17	3
Inbursa	0.50	0.50	0.027	0.084	0.67	1.54	1
Metlife	1.23	0.25	0.000	0.001	0.69	2.67	6
IXE	1.10	0.35	0.000	0.000	0.72	2.42	5
XXI	1.30	0.20	0.041	0.065	0.79	2.89	7
Banamex	1.70	0.00	0.244	0.199	0.80	3.49	12
ING	1.68	0.00	0.085	0.089	0.86	3.44	10
Santander	1.60	0.70	0.117	0.086	0.87	4.01	15
Bancomer	1.68	0.00	0.148	0.226	0.89	3.40	9
Principal	1.60	0.35	0.074	0.039	0.89	3.48	11
HSBC	1.60	0.40	0.042	0.037	1.00	3.67	14
Banorte Generali	1.40	0.50	0.096	0.061	1.07	3.40	8
Profuturo	1.67	0.60	0.122	0.107	1.14	3.64	13

The share of assets is June 2005 is estimated using affiliates' account balances in June 2006 and the afore they were afilliated with in June of 2005. All other statistics are from June 2005. Stataistics are based on a 0.5% random sample of account holders.

TABLE 3: AFORE CHOICE BEFORE AND AFTER INTRODUCTION OF 1 YEAR EQUIVALENT FEE INDEX

	January 2004 - June 2005			July 2005 - December 2006		
	25th Pctl.	Median	75th Pctl.	25th Pctl.	Median	75th Pctl.
Changes in Fees (New Afore – Old Afore)						
Change in Afore CEF Rank	-3	2	5	-6	-2	3
Change in Flow Fee	-0.12	0.01	0.30	-0.34	-0.08	0.09
Change in Balance Fee	-0.21	0.05	0.40	-0.15	0.00	0.15
Remaining Potential Fee Gain (Cheapest Afore - New Afore)						
Remaining Flow Fee Gain	-1.17	-1.10	-0.90	-0.92	-0.77	-0.50
Remaining Balance Fee Gain	-0.60	-0.40	0.00	-0.35	-0.25	-0.15
Changes in Costs Measures in Days of Earnings (New Afore – Old Afore)						
Change in Total Cost Measure	-49.29	16.43	127.57	-82.10	-7.99	47.77
Change in Predicted Cost Measure	-50.01	17.70	128.64	-83.80	-8.78	49.53
Change in Myopic Cost Measure	-44.28	7.51	121.26	-80.17	-3.84	40.70
Remaining Potential Cost Savings in Days of Earnings (Cheapest Afore - New Afore)						
Remaining Total Cost Measure	-256.89	-124.80	-51.03	-214.56	-98.83	-32.43
Remaining Total Predicted Cost Measure	-254.98	-125.55	-52.34	-215.04	-101.75	-35.11
Remaining Total Myopic Cost Measure	-269.38	-115.93	-34.56	-230.35	-88.27	-23.08
N	278,348	278,348	278,348	489,993	489,993	489,993

Notes: Administrative data on account movements between Afores from January 2004 thorugh December 2006. CEF ranking is based on CEF25 from January 2004 through June 2005, and CEF 1 from July 2005 through Decmeber 2006.

TABLE 4: AFORE CHOICE BEFORE AND AFTER FEE INDEX INTERVENTION BY EMPLOYMENT SUBGROUPS

	Always Formally Employed				Never Formally Employed			
	Pre June 2005		Post June 2005		Pre June 2005		Post June 2005	
	Mean	StDev	Mean	StDev	Mean	StDev	Mean	StDev
Changes in Fees (New Afore – Old Afore)								
Change in Afore CEF Rank	0.63	5.59	-2.25	6.31	-0.29	5.46	-1.65	6.70
Change in Flow Fee	-0.04	0.53	-0.15	0.44	-0.10	0.61	-0.11	0.46
Change in Balance Fee	0.05	0.42	0.01	0.28	0.02	0.38	-0.01	0.31
Remaining Potential Fee Gain (Cheapest Afore - New Afore)								
Remaining Flow Fee Gain	-0.92	0.37	-0.67	0.35	-0.83	0.42	-0.69	0.35
Remaining Balance Fee Gain	-0.34	0.27	-0.28	0.18	-0.33	0.23	-0.28	0.20
Changes in Costs Measures in Days of Earnings (New Afore – Old Afore)								
Change in Total Cost Measure	30.14	225.96	-27.20	178.90	5.35	46.91	-0.76	46.72
Change in Predicted Cost Measure	30.28	227.23	-27.46	179.92	6.17	54.87	-1.40	52.72
Change in Myopic Cost Measure	28.30	226.40	-27.97	185.23	5.35	46.91	-0.76	46.72
Remaining Potential Cost Savings in Days of Earnings (Cheapest Afore - New Afore)								
Remaining Total Cost Measure	-200.00	201.22	-162.02	169.45	-25.02	47.03	-24.93	44.69
Remaining Total Predicted Cost Measure	-200.88	203.12	-162.81	170.83	-30.81	52.30	-29.35	50.29
Remaining Total Myopic Cost Measure	-194.71	210.15	-167.96	178.93	-25.02	47.03	-24.93	44.69
N	117,165		191,528		5,923		14,497	

Notes: Administrative data on account movements between Afores from January 2004 thorugh December 2006. CEF ranking is based on CEF25 from January 2004 through June 2005, and CEF 1 from July 2005 through Decmeber 2006.

TABLE 5: AFORE CHOICE BEFORE AND AFTER FOR EMPLOYMENT AND WAGE SUBGROUPS

	Always Formally Employed and High Earner				Never Formally Employed and High Earner			
	Pre June 2005		Post June 2005		Pre June 2005		Post June 2005	
	Mean	StDev	Mean	StDev	Mean	StDev	Mean	StDev
Changes in Fees (New Afore – Old Afore)								
Change in Afore CEF Rank	0.23	5.35	-2.48	6.16	-1.07	5.46	-2.65	6.91
Change in Flow Fee	-0.08	0.51	-0.16	0.44	-0.21	0.66	-0.19	0.50
Change in Balance Fee	0.04	0.43	0.00	0.28	0.02	0.39	0.00	0.31
Remaining Potential Fee Gain (Cheapest Afore - New Afore)								
Remaining Flow Fee Gain	-0.91	0.38	-0.65	0.35	-0.72	0.48	-0.60	0.39
Remaining Balance Fee Gain	-0.33	0.27	-0.27	0.18	-0.34	0.23	-0.29	0.21
Changes in Costs Measures in Days of Earnings (New Afore – Old Afore)								
Change in Total Cost Measure	10.72	175.85	-31.51	141.34	5.41	34.56	0.52	30.82
Change in Predicted Cost Measure	10.79	176.74	-31.72	142.40	5.37	35.92	0.22	32.11
Change in Myopic Cost Measure	9.07	169.80	-32.77	148.85	5.41	34.56	0.52	30.82
Remaining Potential Cost Savings in Days of Earnings (Cheapest Afore - New Afore)								
Remaining Total Cost Measure	-169.74	153.61	-137.06	132.56	-20.57	32.15	-21.02	29.25
Remaining Total Predicted Cost Measure	-170.38	155.31	-137.79	133.90	-21.41	32.15	-21.02	29.25
Remaining Total Myopic Cost Measure	-162.02	151.74	-143.76	143.08	-20.57	30.99	-19.98	27.80
N	80,132		130,396		1,897		4,078	

Notes: Administrative data on account movements between Afores from January 2004 thorugh December 2006. CEF ranking is based on CEF25 from January 2004 through June 2005, and CEF 1 from July 2005 through Decmeber 2006.

TABLE 6: AFORE CHOICE BEFORE AND AFTER INDEX INTERVENTION BY AGE AND EXPERIENCE

	Under 30 yrs old		Between 30 and 40		Over 40 yrs old		Over 7 yrs in system		Under 3 yrs in system	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
Changes in Fees (New Afore – Old Afore)										
Change in Afore CEF Rank	1.58	-1.50	0.80	-2.24	0.92	-2.26	-0.35	-2.48	3.26	0.56
Change in Flow Fee	0.08	-0.10	-0.02	-0.15	-0.01	-0.15	-0.18	-0.16	0.56	0.00
Change in Balance Fee	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.00	0.05	0.03
Remaining Potential Fee Gain (Cheapest Afore - New Afore)										
Remaining Flow Fee Gain	-0.98	-0.74	-0.91	-0.65	-0.89	-0.63	-0.84	-0.63	-1.05	-0.81
Remaining Balance Fee Gain	-0.34	-0.29	-0.35	-0.28	-0.36	-0.28	-0.33	-0.28	-0.37	-0.29
Changes in Costs Measures in Days of Earnings (New Afore – Old Afore)										
Change in Total Cost Measure	59.52	-15.94	14.91	-19.73	5.27	-9.35	-0.23	-25.32	93.87	19.61
Change in Predicted Cost Measure	59.02	-16.54	14.55	-20.01	5.07	-20.01	-0.52	-26.00	93.06	19.93
Change in Myopic Cost Measure	67.24	-16.42	17.44	-19.38	6.58	-8.82	-0.16	-25.63	106.67	22.35
Remaining Potential Cost Savings in Days of Earnings (Cheapest Afore - New Afore)										
Remaining Total Cost Measure	-243.03	-201.12	-106.91	-104.04	-40.89	-39.52	-142.47	-127.03	-220.91	-195.34
Remaining Total Predicted Cost Measure	-243.14	-203.62	-107.00	-104.49	-41.01	-39.68	-143.39	-129.01	-219.92	-193.10
Remaining Total Myopic Cost Measure	-260.42	-213.79	-113.76	-104.01	-44.25	-39.25	-147.17	-128.12	-242.23	-211.79
N	181,175	291,616	61,745	123,729	35,428	74,648	82,032	247,168	123,922	96,496

TABLE 7: SENSITIVITY OF CEF1 VERSUS CEF 25 TO FLOW AND BALANCE FEES

Afore	Derivative of 25-year CEF w.r.t. balance fee	Derivative of 25-year CEF w.r.t. flow fee	Derivative of 1-year CEF w.r.t. balance fee	Derivative of 1-year CEF w.r.t. flow fee
Actinver	1.014	0.539	0.991	2.073
Azteca	1.016	0.545	0.990	2.073
Banamex	1.035	0.583	0.983	2.075
Bancomer	1.046	0.593	0.983	2.075
Banorte Generali	1.000	0.571	0.985	2.069
HSBC	1.003	0.571	0.984	2.070
ING	1.044	0.590	0.983	2.075
IXE	1.003	0.544	0.990	2.071
Inbursa	1.001	0.527	0.995	2.069
Invercap	1.014	0.544	0.990	2.073
Metlife	1.008	0.549	0.988	2.072
Principal	1.002	0.563	0.984	2.071
Profuturo	1.000	0.570	0.985	2.068
Santander	0.928	0.532	0.983	2.067
XXI	1.025	0.565	0.987	2.073
Total	1.009	0.559	0.987	2.072

All statistics are from June 2005.

TABLE 8: CHANGES IN FEE STRUCTRE BETWEEN JUNE 2005 AND DECEMBER 2007

Afore	Flow Fee in June 2005	Change in Flow Fee	Balance Fee in June 2005	Change in Balance Fee	%Change in CEF 1
Actinver	1.03	-0.02	0.20	0.00	-11.27%
Azteca	1.10	-0.20	0.15	0.25	-12.16%
Banamex	1.70	-0.95	0.00	1.48	-34.09%
Bancomer	1.68	-0.48	0.00	0.50	-41.82%
Banorte Generali	1.40	-0.70	0.50	0.64	-49.99%
HSBC	1.60	-0.85	0.40	0.80	-39.13%
Inbursa	0.50	0.00	0.50	0.00	-0.01%
ING	1.68	-0.98	0.00	1.45	-48.54%
Invercap	1.03	-0.23	0.20	0.16	-11.80%
Ixe	1.10	-0.34	0.35	-0.10	-24.63%
Metlife	1.23	-0.03	0.25	0.46	-19.34%
Principal	1.60	0.00	0.35	0.00	-0.01%
Profuturo GNP	1.67	-0.07	0.60	0.60	-35.32%
Santander	1.60	-0.90	0.70	0.75	-55.78%
XXI	1.30	-0.08	0.20	0.06	-7.98%

Notes: Fee structures downloadable from www.consar.gob.mx

TABLE 9: ESTIMATED MEAN ELASTICITIES FOR AFORE SWITCHERS
PRE AND POST REFORM

	Elasticity w.r.t CEF		Elasticity w.r.t. Management Cost	
Afore	Pre	Post	Pre	Post
Actinver	-0.211	-0.906	0.003	0.051
Afirme	--	-0.693	--	0.056
Azteca	-0.211	-0.950	-0.001	0.046
Banamex	-0.245	-1.293	-0.026	0.070
Bancomer	-0.249	-1.209	-0.003	0.085
Banorte Generali	-0.357	-1.237	0.000	0.088
Coppel	--	-1.006	--	0.073
HSBC	-0.336	-1.336	-0.013	0.079
Inbursa	-0.217	-0.616	0.019	0.066
ING	-0.243	-1.370	0.003	0.085
Invercap	-0.257	-0.959	-0.036	0.052
IXE	-0.266	-1.096	-0.001	0.075
MetLife	-0.282	-1.194	-0.039	0.067
Principal	-0.286	-1.484	0.006	0.090
Profuturo GNP	-0.349	-1.289	0.006	0.100
Santander	-0.276	-1.448	0.026	0.110
XXI	-0.261	-1.290	-0.002	0.067
N	2,732,799	5,824,526	2,732,799	5,824,526

Notes:

TABLE 10: BEST RESPONSE TO INFORMATION MANDATE AND PREFERENCE CHANGES

Afore	June 2005 market, Old Preferences and CEF25		June 2005 market, New Preferences, CEF1	
	Best Response Flow fee	Best Response Balance Fee	Best Response Flow fee	Best Response Balance Fee
Actinver	1.3	0	0	2.0
Azteca	1.2	0	0	2.2
Banamex	1.6	0	0	3.4
Bancomer	1.6	0	0	3.1
Banorte Generali	1.6	0	0	3.0
HSBC	1.9	0	0	3.0
Inbursa	1.3	0	0	1.5
ING	1.4	0	0	2.6
Principal	1.7	0	0	2.7
Profuturo GNP	1.9	0	0	3.1
Santander	1.5	0	0	3.4
XXI	1.5	0	0	2.8

Notes:

TABLE 11: IMPACT OF POLICY AND FIRM RESPONSE ON EXPECTED MANAGEMENT COSTS BY DEMOGRAPHICS

	Percent Change in Expected Costs (Cost at Dec. '07 fees - Cost at June '05 fees)/ (Cost at June '05 fees)	
Wage quartile (among account movers)	Females	Males
1	43.5%	50.2%
2	-16.1%	-13.9%
3	-18.4%	-19.6%
4	-21.8%	-21.7%
Overall		-13.5%

Figure 1: Movement of Account Switching Before and After Information Reform by Change in Cost and CEF

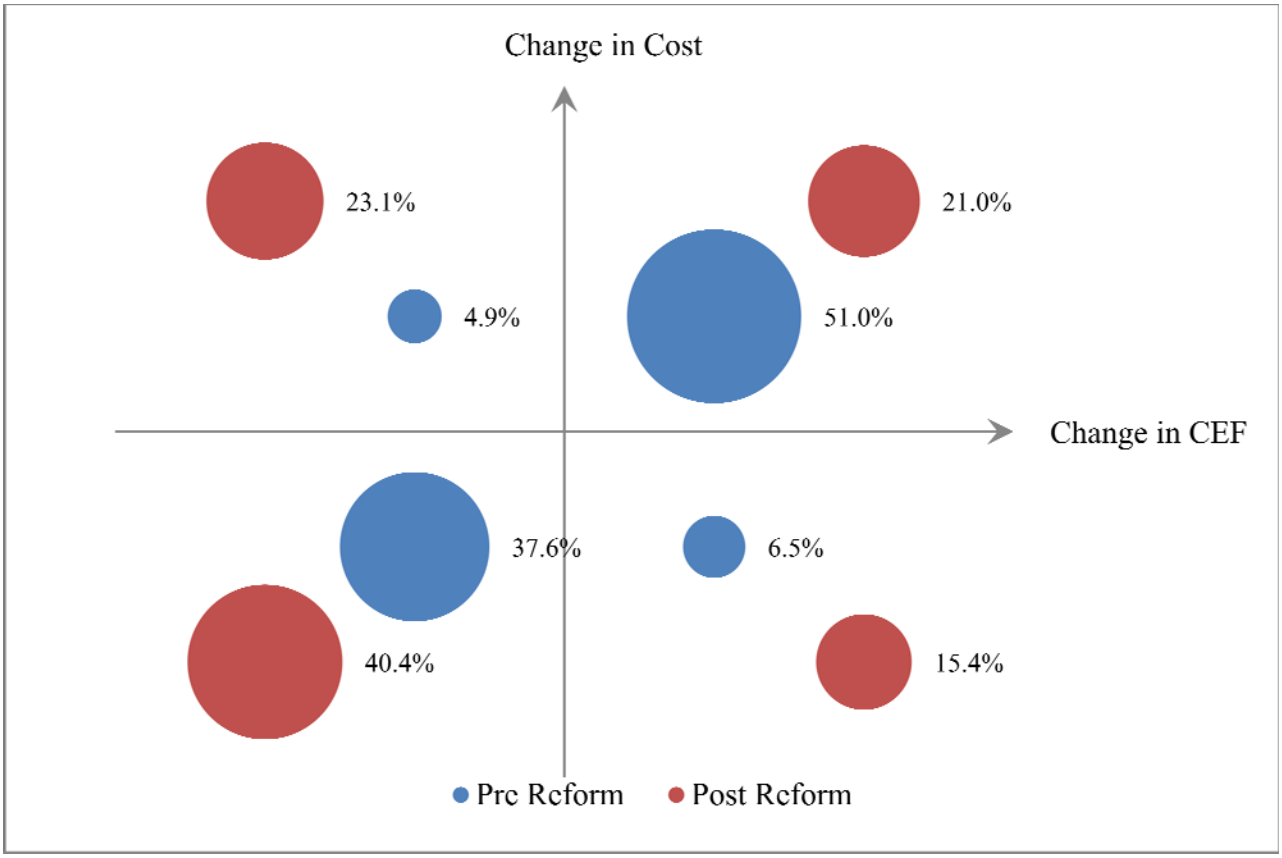


Figure 2a. Afore CEF, June 2004 - June 2007

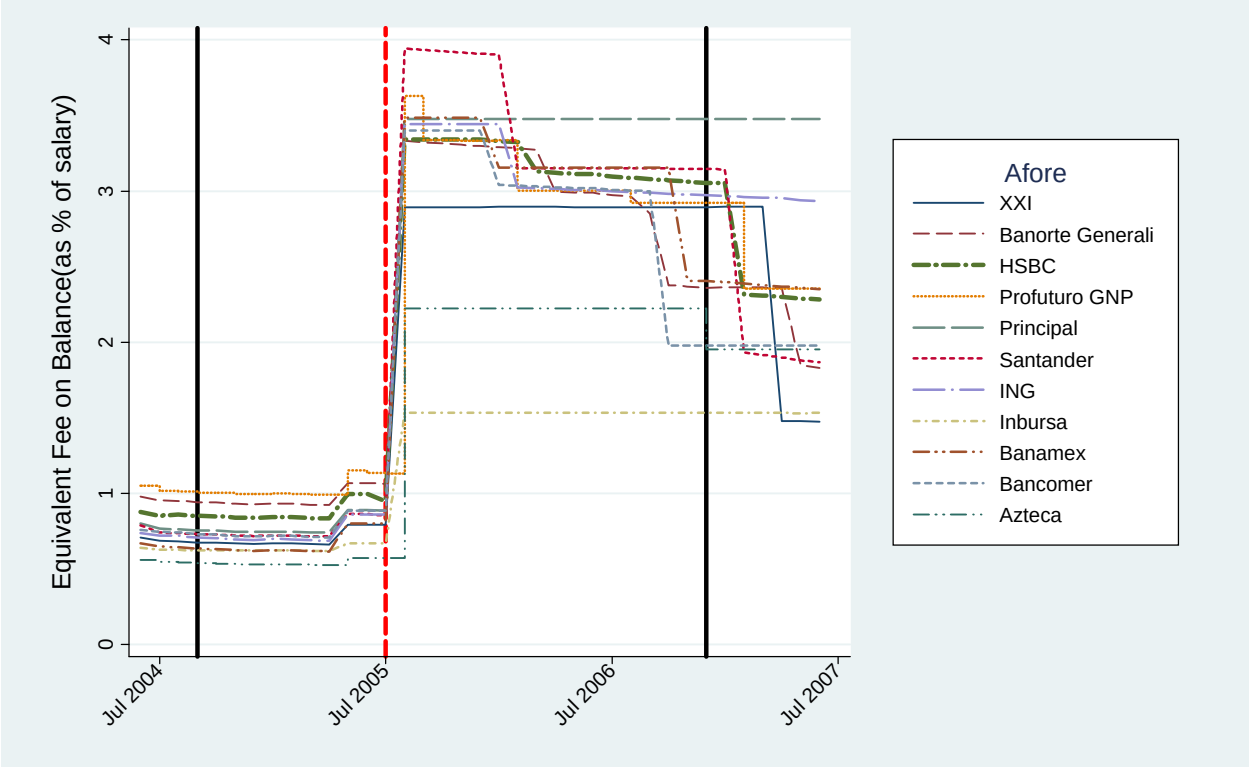


Figure 2b. Afore CEF, June 2004 - June 2007

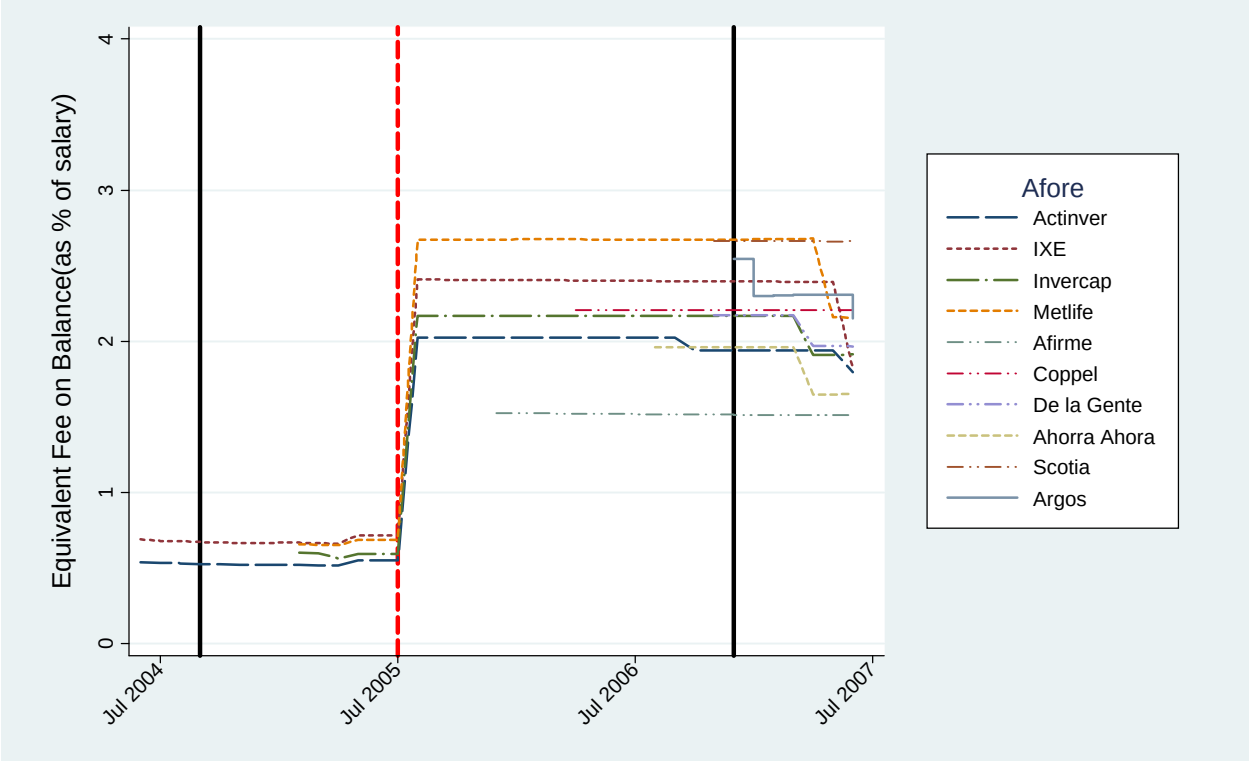


Figure 3: Flow Fee Changes

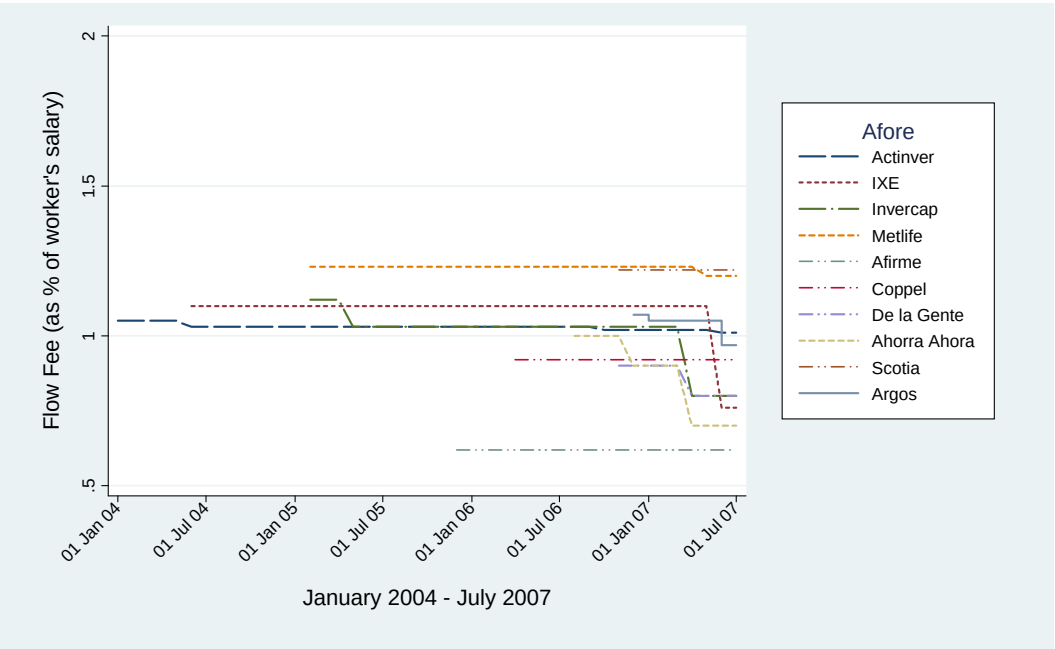
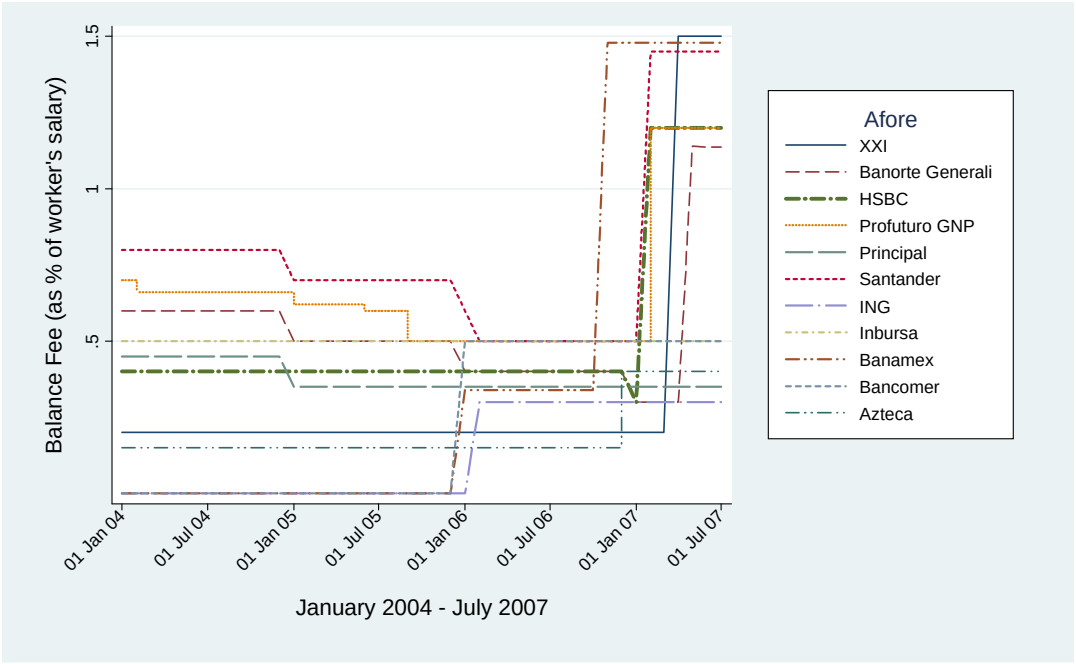
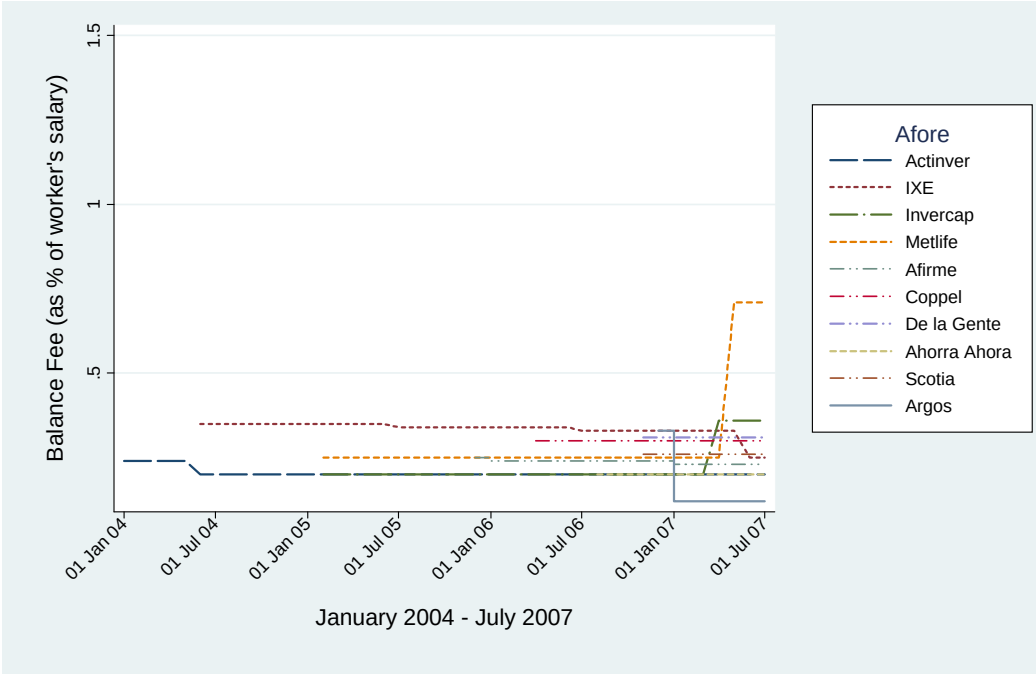


Figure 4: Balance Fee Changes





PROJECTION BIAS IN THE CAR AND HOUSING MARKETS

Meghan R. Busse
Devin G. Pope
Jaren C. Pope
Jorge Silva-Risso*

June 2012

Abstract

Projection bias is the tendency to overpredict the degree to which one's future tastes will resemble one's current tastes. We test for evidence of projection bias in two of the largest and most important consumer markets – the car and housing markets. Using data for more than forty million vehicle transactions and four million housing purchases, we explore the impact of the weather on purchasing decisions. We find that the choice to purchase a convertible, a 4-wheel drive, or a vehicle that is black in color is highly dependent on the weather at the time of purchase in a way that is inconsistent with classical utility theory. Similarly, we find that the hedonic value that a swimming pool and that central air add to a house is higher when the house goes under contract in the summertime compared to the wintertime.

*Busse: Kellogg School of Management, Northwestern University and NBER; D. Pope: Booth School of Business, University of Chicago and NBER; J. Pope: Department of Economics, Brigham Young University; Silva-Risso: School of Business Administration, UC Riverside. We are grateful to Chris Bruegge and Ezra Karger for valuable research assistance. We also thank Stefano DellaVigna, Emir Kamenica, Ulrike Malmendier, Ted O'Donoghue, Loren Pope, Mathew Rabin, Dick Thaler, and seminar participants at the Behavioral Economics Annual Meeting, Columbia University, Cornell University, Harvard Business School, MIT, Stanford University, UC Berkeley, UC Los Angeles, UC San Diego, UC Santa Barbara, University of Chicago, University of Illinois, University of Zurich, and Yale University for helpful suggestions.

Many decisions that people make require them to predict their future utility. For example, choosing a job, deciding where to live, planning a vacation, buying a car, deciding whether to have a baby, and purchasing a home are all important life decisions that require predicting future utility across a variety of choice dimensions. The standard economic model assumes that an individual's prediction of future utility will, on average, match his or her realized utility. Evidence from psychology, however, suggests that individuals may be systematically biased when predicting future utility. A specific bias that has received considerable attention is projection bias (Loewenstein, O'Donoghue, and Rabin, 2003). Projection bias refers to the tendency of individuals to overpredict the degree to which their future tastes will resemble their current tastes. For example, the popular adage "never shop on an empty stomach" is a caution against projection bias: consumers are likely to overpredict the degree to which their future selves will appreciate the purchases that their current selves crave. While projection bias has intuitive appeal for situations such as shopping while hungry, an open question is whether this bias influences important life decisions for which people are likely to have strong motivations to make a good decision.

In this paper, we test for projection bias in two high-stakes environments: the purchases of vehicles and houses. Vehicles and houses are durable goods. When consumers purchase durable goods, they must predict at the time of purchase how much they will value consuming these goods in the future, including the enjoyment they will experience in a variety of future states of the world. Projection bias suggests that consumers may mistakenly purchase a vehicle or a house that has a high utility at the time of purchase, but whose utility will not be as high in other states of the world that the consumer will experience while owning the vehicle or house. We test the extent to which weather variation at the time of purchase can cause consumers to overweigh the value that they place on certain vehicle and housing characteristics. Projection bias predicts that consumers will overvalue warm-weather vehicle types and housing characteristics (e.g. convertibles and swimming

pools) when the weather is warm at the time of purchase and overvalue cold-weather vehicle types (e.g. 4-wheel drive vehicles) when the weather is cold and snowy at the time of purchase.

We begin by exploring these hypotheses in the car market using transaction-level data for more than forty million transactions of new and used vehicles from dealerships around the U.S. We find that the sales of convertibles, 4-wheel drives, and vehicles that are black in color are highly influenced by idiosyncratic variation in temperature, cloud cover, and snowfall. We show that for convertibles, weather that is warmer and skies that are clearer than seasonal averages lead to a higher number of sales. Controlling for seasonal sales patterns, our estimates suggest that a location that experiences a mean temperature that is 20 degrees higher than normal will experience a 0.22 percentage point increase in the percentage of total vehicles sold that are convertibles. Given a base rate of 2.6% of vehicles sold that are convertibles, this represents an 8.5% increase in the fraction of convertible cars sold. We find large and significant effects both in the spring and in the fall (e.g. an abnormally warm week in November increases the fraction of vehicles sold that are convertibles). Importantly, we also show that abnormally warm weather does not impact convertible sales when the temperature is already high (when average daily high temperature is already more than about 80 degrees Fahrenheit). Purchases of 4-wheel drive vehicles are also very responsive to abnormal weather variation—particularly snowfall. Our results suggest that a snow storm of approximately 10 inches will increase the fraction of vehicles sold that have 4-wheel drive by about 2 percentage points over the next 2-3 weeks (an approximately 6% increase over the base rate of 33.5%). This effect is robust to using an event study design that uses large storms as events. Black vehicles are less likely to be purchased when the weather is warm and sunny. A 20-degree increase in temperature leads to a 2.1% (0.26 percentage point) reduction in the fraction of vehicles sold that are black, compared to a 12.6% baseline percentage. Moving from overcast to completely clear weather reduces the sales of black vehicles by 5.6% (0.71 percentage points).

The data allow us to rule out several alternative explanations for these findings. For example, a distributive-lag model indicates that the increase in convertible sales and most of the increase in 4-wheel drive sales due to abnormal weather cannot be explained by short-run substitutions in vehicle purchases from one week to the next (a "harvesting effect"). We also present evidence that learning about a vehicle during a test drive (which for a convertible may be easier to do on a warm day) is unlikely to explain the results we find. In particular, cloud cover (which does not limit the ability to test drive a vehicle as temperature might) has a large impact on sales. Furthermore, individuals who previously owned a convertible and thus have less to learn about their value for convertible attributes are also affected by idiosyncratic weather conditions. Finally, we look at the impact of the weather at the time of vehicle purchase on the probability that a vehicle is traded in quickly for a different vehicle. This analysis, which uses unique vehicle identifiers to follow vehicles over time in our data, suggests that a vehicle is more likely to be returned quickly when purchased on a day with abnormal weather—evidence in favor of projection bias.

The second part of the paper turns to identifying projection bias in the housing market using a repeat-sales methodology for over four million housing transactions. This methodology allows us to estimate the value that certain house characteristics (e.g. a swimming pool or central air) have at different times of the year by looking at two different sales for a single house, while also controlling for variation in overall housing trends across time and space. We find evidence that a swimming pool adds more value to a house that goes under contract in the summertime than it adds to the same house that goes under contract in the wintertime. Specifically, a house with a swimming pool that goes under contract in the summertime sells for an average of 0.4 percentage points more than the same house when it goes under contract in the wintertime. Given the average value of homes with swimming pools in our dataset, this effect suggests a swing in value of approximately $1600 between summer and winter contract dates.

This result is robust to a variety of different specifications and subsamples of the data. Our within-house identification strategy helps us to rule out concerns about unobserved housing characteristics that are correlated with houses that have swimming pools or with the type of people who buy and sell houses with swimming pools. Our fixed-effects framework also allows us to control for seasonal patterns in houses overall in order to identify the interaction between seasonal weather and houses with swimming pools. We also discuss and rule out the possibility that a home with a swimming pool may be worth more due to immediate utility gains (during the season of purchase). Finally, we provide the results for three other housing characteristics whose value may fluctuate across seasons—central air, lot size, and fireplaces. We also find evidence that the value of central air is higher when a home sells in the summertime. However, we find no evidence that the hedonic value of lot size or fireplaces vary with seasonal temperature and discuss likely explanations for this finding.

Our findings are significant for several reasons. First, the car and housing markets in and of themselves are large and important. Identifying, and potentially correcting, systematic errors in these markets can have valuable welfare implications. Perhaps more importantly, our results suggest that projection bias may be prevalent in other important decisions (getting married, choosing a job, etc.) that are similarly distinguished by having large stakes, state-dependent utility, and low-frequency decision-making.

Our paper is related to a growing literature that uses field data to test models from behavioral economics (see DellaVigna (2009) for a review). More specifically, our paper relates to a small literature that empirically explores projection bias in field settings (Read & van Leeuven, 1998; Conlin, O'Donoghue, & Vogelsang, 2007; Simonsohn, 2010).[1] Our paper is most similar to the work

[1] In the psychology literature, the type of projection bias that we explore in this paper is most closely related to the work on hot/cold empathy gaps and visceral states (see for example, Nisbett and Kanouse (1968), Loewenstein (1996), Loewenstein, Nagin, & Paternoster (1997), Van Boven & Loewenstein (2003), Nordgren, van der Pligt, and van

of Conlin, O'Donoghue, and Vogelsang (2007) who test for projection bias in catalog orders. They convincingly show that decisions to purchase cold-weather items are overinfluenced by the weather at the time of purchase. Specifically, they find that if the temperature at the time of a purchase is 30 degrees lower, consumers are 0.57 percentage points more likely to return the item (3.95%). Our paper complements this earlier work. We extend the existing research by providing evidence of projection bias in two markets of even greater economic importance. The richness of our data allows us to explore not only how projection bias impacts sales volume, but also whether it has an impact on prices.

The paper proceeds as follows. Section I provides a simple, conceptual framework for projection bias following Loewenstein, O'Donoghue, & Rabin (2003). Section II explores the data, empirical strategy, and results for the car market. Section III describes the data, empirical strategy, and results for the housing market. Section IV provides a conclusion along with a brief discussion of the broader implication of our findings.

I. Conceptual Framework

In this section we describe how projection bias may influence durable goods purchases, following the framework of Loewenstein, O'Donoghue, & Rabin (2003). To begin, suppose that a person has state-dependent utility such that her instantaneous utility of consumption, c, in state, s, can be represented as $u(c,s)$. Furthermore, consider an individual who is currently in state s' who is attempting to predict her future instantaneous utility of consumption, c, in state s: $\tilde{u}(c,s|s')$. An accurate prediction would be represented by $\tilde{u}(c,s|s') = u(c,s)$.

Loewenstein, O'Donoghue, & Rabin (2003) argue that projection bias causes agents' predictions about future utility to be unduly influenced by the state they are in at the time of the prediction. Specifically, an individual exhibits projection bias if

Harreveld (2006, 2007). Loewenstein and Schkade (1999) provide a useful review of the psychological evidence for projection bias.

(1) $$\tilde{u}(c, s|s') = (1-\alpha)u(c, s) + \alpha\big(u(c, s')\big),$$

where α is a number between 0 and 1. If $\alpha = 0$, then the individual accurately predicts her future preferences, whereas if $\alpha > 0$, an individual perceives her future utility to reflect a combination of her true future utility along with the utility that consumption c would provide in her current state s'.

This simple model of projection bias can be extended to an intertemporal-choice framework. Consider, for example, the instantaneous utility that a person receives in time t from purchasing a convertible in time t $(conv_t)$ and owning it until period T. Her true utility can be represented by

(2) $$U^t(conv_t, \ldots, conv_T) = \sum_{\tau=t}^{T} \delta^\tau u(conv_\tau, s_\tau),$$

where $0 \leq \delta \leq 1$ is her standard discount factor. Once again, following Loewenstein, O'Donoghue, & Rabin (2003), a person with projection bias perceives her intertemporal utility to be

(3) $$\tilde{U}^t(conv_t, \ldots, conv_T|s_t) = \sum_{\tau=t}^{T} \delta^\tau \tilde{u}(conv_\tau, s_\tau|s_t),$$

where $\tilde{u}$ represents the perceived instantaneous utility described by Equation (1).

This framework illustrates that an individual's perceived intertemporal utility of purchasing a convertible at time t, $\tilde{U}^t$, is overly influenced by s_t. Specifically, we would predict that when s_t is a very good state of the world for consuming a convertible (warm, sunny weather), an individual has a higher perceived utility of purchasing the convertible than when s_t is a bad state of the world for consuming a convertible (cold, cloudy weather).

A challenge involved with empirically testing for projection bias is that the state at the time of purchase s_t, while unduly influential for agents with projection bias, also matters for agents that do not have projection bias (see Equation (2)). If the number of periods is small, then it would be perfectly reasonable that the current state of the world has an important impact on the decision to buy. For example, in an extreme case, imagine an individual's decision to rent a car for a few days. It would be perfectly reasonable to be more likely to rent a convertible if the weather on the day of rental is nice since the consumption utility from the first period is a large part of the overall consumption utility. The advantage that we have in our paper is that we are focused on the purchases of very durable goods (vehicle and home purchases). For these purchases, we argue that

idiosyncratically warm weather on the day of purchase should have a minimal impact on the probability of purchasing a particular type of vehicle or house since most vehicles and houses are owned for a considerable period of time. In fact, even very high initial discount rates consistent with present-biased preferences of the type described by Laibson (1997) and O'Donoghue and Rabin (1999) cannot easily explain the effect sizes that we find on vehicle purchases. It would be even harder to explain our housing purchase results using present-biased preferences because housing consumption does not occur at the time of the decision (houses go under contract many days before a house is sold).

It is also important to note that weather states are not uncorrelated. In general, there are many warm-weather states that occur sequentially in the summertime followed by many cold-weather states in the wintertime. We would expect an unbiased, rational agent to be more willing to purchase a convertible in the spring than in the fall since a consumer who buys in the spring is likely to experience a string of "good" states of the world starting immediately. A consumer who buys in the fall will have to wait months to experience a similar run of "good" states of the world. Similarly, one might imagine that home buyers would be willing to pay slightly more for a home with a swimming pool when they are moving in at the beginning of the summer (and can use the pool immediately) relative to the amount they would be willing to pay if they moved in after the end of the summer (and would have to wait until next summer to use the pool). Thus, simply finding that people are willing to pay more for a home with a swimming pool or are more likely to buy a convertible when the weather is nice outside could be a response by agents who are accurately predicting their future utility and does not necessarily provide evidence of projection bias.

Our empirical strategies allow us to overcome this identification problem. In the housing market, we overcome this problem by using the fact that the purchase decision of a home (the date the home goes under contract) is made, on average, two months before the closing date. This lag between the decision and move-in dates allows us to distinguish between a rational response to the weather state at the time of purchase and a response by agents with projection bias. Specifically, we find evidence that swimming pools are very highly valued when homes go under contract in August

(the hottest month of the year). While this fits a model of projection bias (since it is the state at the time of the decision that matters), it is not consistent with a more standard model of how people should value a swimming pool since the home buyers will likely move into their homes in October or later (perhaps the worst time from a rational perspective to purchase a house with a swimming pool). In the car market, we utilize idiosyncratic weather shocks to overcome this identification problem. Specifically, we control for the time of year when the vehicle purchase is made and test for the impact of abnormally warm or cold weather on purchase decisions. By controlling for the time of year, this strategy eliminates all seasonal patterns in vehicle purchases (e.g. the value to purchasing a convertible in the spring rather than the fall).

One final note regarding our conceptual framework relates to whether or not individuals correctly anticipate the path of states ($s_t, \ldots s_T$). It is possible that individuals are more likely to predict a greater number of warm-weather states in the future when the current weather is warm relative to when the current weather is cold.[2] Loewenstein, O'Donoghue, & Rabin (2003) assume that individuals correctly anticipate the path of states, but err when predicting the utility that those states, combined with a given consumption, will generate. In practice, these two errors (projection bias of utility and projection bias of states) both lead to similar incorrect predictions of future utility. Thus, it is difficult to separate these two different types of projection bias and our analysis will not attempt to do so. However, there are several reasons to believe that projection bias of states is unlikely to be the underlying mechanism. The first is the prevalence of weather information that is available to people during the time of our study, including their own experience of local weather patterns. It is much harder to find information about future utility than it is to find information about future states. In addition, Conlin, O'Donoghue, & Vogelsang (2007) (who also comment on this question) cite Krueger & Clement (1994) who find that students at Brown University did a reasonable job of estimating temperature levels in Providence for different days of the year.

[2] Some psychological evidence suggests that being in a hot or cold state may make associated states of the world seem more likely in the future (see for example, Risen & Critcher (2011) and Li, Johnson, and Zaval (2011)).

II. Car Market

Data and Empirical Strategy. The data used in our analysis contain information about automobile transactions from a sample of about 20% of all new car dealerships in the U.S. from January 1, 2001 to December 31, 2008. The data were collected by a major market research firm, and include every new and used vehicle transaction that occurred at the dealers in the sample. For each transaction, we observe the date and location of the purchase, information about the vehicle purchased, and the price paid for the vehicle. Our locations are defined by Nielsen Designated Market Areas (DMAs), which divide the U.S. into approximately 200 areas. DMAs are defined to correspond to media markets, which means that DMAs corresponding to major cities will have higher populations than DMAs in more rural regions. Examples of DMAs in our data include Phoenix, Arizona; Tulsa, Oklahoma; Lansing, Michigan; and Billings, Montana.[3]

We will add to these data information about local weather. The weather data were collected by first using wolframalpha.com to find the weather station nearest to the principal city in each DMA. Weather data themselves were obtained for each weather station from Mathematica's WeatherData compilation.[4] Data were collected on temperature, precipitation, precipitation type, and cloud cover. Temperature is measured as the simple average of the seven daily high temperatures in the week, measured in degrees Farenheit. Precipitation is measured as the cumulative liquidized inches over the course of the week. If the only precipitation type reported in the week is rain, we classify the precipitation as rainfall (measured in inches). If the only precipitation type reported during the week is snow, we classify the precipitation as snowfall (measured in liquidized inches). If both rain and snow are reported during the week, we classify the precipitation as slushfall (measured in liquidized inches). Cloud cover is a simple average of the seven daily measures of the fraction of the sky covered by clouds.

The data indicate that vehicle transactions occur all year round, but are most common during the summer months. Of primary interest in this paper is the seasonal trend in convertible and 4-

[3] A list of all the DMAs covered by our data is available from the authors.
[4] If the weather station did not have weather data available for at least 90% of the 4745 daily observations between 1997 and 2010, data for the second- or third-closest weather station was used for that DMA. (There are 21 DMAs that use data from the second-closest station, and 6 that use data from the third-closest station.)

wheel drive purchases. In Panel A of Figure 1, we illustrate the percentage of total vehicle transactions that were convertibles by month of the year. Overall, convertibles make up between 1.5 and 3% of total vehicles purchased. The data show a strong seasonal pattern in which the percentage of vehicles sold that are convertibles is highest in the early spring. For seven out of the eight years, the percentage of vehicles purchased that are convertibles peaks in April. While springtime is the most popular time to buy a convertible, the percentage of vehicles sold that are convertibles is still relatively large in the winter months. The annual winter troughs in percentage of vehicles sold that are convertibles are well over half the magnitude of the corresponding spring peaks. These seasonal differences in convertible purchases are consistent with the standard model of state-dependent preferences discussed in the conceptual framework section: consumers do seem to take into consideration the season of the year when making convertible purchases since those first few months of consumption in the warm-weather state will likely increase total discounted utility for spring buyers relative to fall buyers.

Similarly, Panel B of Figure 1 illustrates the percentage of total vehicle transactions that were 4-wheel drive vehicles by month of the year. 4-wheel drive transactions range between 20% and 35% of total vehicle transactions. Panel B shows a seasonal pattern in which 4-wheel drive vehicles are particularly popular in the early winter months (purchases usually peak in December).[5] As was the case for convertibles, this is not yet strong evidence for projection bias since a standard model of state-dependent preferences would predict that the discounted utility of a 4-wheel drive is highest at the beginning of the winter.

We expect there to be a large amount of heterogeneity in the seasonal differences shown in Figure 1 depending on the geographic location of the dealership. To illustrate this heterogeneity, we perform a simple cut of the data by dividing DMAs into two groups: DMAs with above- and DMAs with below-median monthly temperature variation.[6] Figure 2, like Figure 1, displays month-to-month sales of convertibles (Panel A) and 4-wheel drive vehicles (Panel B) as a percentage of total

[5] There is a mid-summer peak in 2005 which arose from record sales during GM, Chrysler, and Ford's employee discount pricing promotions. (Busse, Simester, and Zettelmeyer (2010) describe the effect of these promotions.)
[6] For each DMA, we calculate the variance of month-by-month average temperature data. DMAs are then classified as above the median if their temperature variance is larger than the median temperature variance in the sample.

vehicles sold, but does so separately by the variable temperature areas (e.g. Chicago) and the non-variable temperature areas (e.g. Miami). Perhaps surprisingly, Panel A shows that the overall percentage of convertibles purchased in these two types of DMAs is not too different. However, it is clear that the amount of seasonal variation is higher in the variable-temperature DMAs. Panel B shows that there is a large level difference in the percentage of 4-wheel drive vehicles purchased in the two types of DMAs and once again the variable temperature areas appear to have a more pronounced seasonal pattern.

Our identification strategy involves testing whether abnormal weather conditions (controlling for time of year in order to eliminate seasonal purchasing patterns) are correlated with abnormally high or low sales volume of convertible and 4-wheel drive vehicles. To do this, we collapse the data to the DMA-week level.[7] After collapsing, we create variables that represent the percentage of total vehicles sold in each DMA-week that were convertibles and that were 4-wheel drive vehicles. Weekly weather data at the DMA level are also merged in. These data will allow us to test whether abnormal weather leads to abnormally high or low levels of convertible and 4-wheel drive purchases. Note that our estimates will identify the effect of weather on the equilibrium sales of vehicles of different types. In other words, we will estimate not only the effect of weather on vehicle demand, but also the effect of any actions dealers take in response to their perception of increased demand for certain types of vehicles under particular weather conditions. Of course, if there is a supply effect, that is evidence that dealers believe buyers are influenced by projection bias, and respond accordingly. Our estimates identify the combined effect of changes in consumers' behavior and dealers' responses to those changes.

We proceed by first presenting the results for convertibles followed by the results for 4-wheel drive vehicles. Vehicles have other characteristics whose value might be weather related (sun roofs, air conditioning, snow tires, etc.). Many of these characteristics are either unobservable to us, or do

[7] Alternatively, the data can be collapsed to the day level. We choose to do most of the analysis at the week level for three primary reasons. First, while the data contain an exact day of purchase, the paperwork may be signed and dated later than the actual date the deal was made. Thus, using day-to-day level variation is noisier than variation at the week level. Second, week-level data largely eliminates the need to worry about weekday/weekend effects as well as holidays and other events that can cause abnormal sales volume. Third, many weather events (e.g. snow storms) occur across multiple days making a weekly analysis more appropriate.

not vary significantly in the data. However, in a later section we will consider the effect of weather on the sales of black vehicles.

Baseline Convertible Results. We begin the analysis by using two DMAs (Chicago and Miami) as examples of the effects that we find. Panel A of Figure 3 plots the percentage of all vehicles sold in Chicago that were convertibles for each week between 2001 and 2008. As expected given the temperature variation that exists in Chicago, we see a strong seasonal pattern in which convertible sales range from approximately 1.5% of total vehicles sold in the wintertime to 3-4% of total vehicles sold in the spring. In accordance with our empirical strategy outlined above, we want to obtain a measure of abnormal convertible sales. To do this, we regress the weekly convertible percentage of total vehicles sold in Chicago on year fixed effects and week-of-the-year fixed effects. The residuals from this regression, which range from approximately -0.75% to 1% are plotted in Panel B of Figure 3. A week with a 0.5% residual is a week in which the convertible percentage of total vehicles sold was 0.5 percentage points higher than our regression predicted for that week of the year. Figure 4 illustrates the seasonal pattern of temperature by week in Chicago. Panel A of Figure 4 shows the average daily temperature for each week. Panel B, which once again nets out year and week-of-the-year effects, illustrates that any given week in Chicago may be up to 20 degrees Fahrenheit hotter or 20 degrees Fahrenheit colder than would be predicted by average seasonal patterns in the data.

To test for projection bias, we want to know whether the abnormal convertible sales illustrated in Panel B of Figure 3 are positively correlated with the abnormal temperature values in Panel B of Figure 4. We find that these residuals are positively and statistically significantly correlated (correlation coefficient = 0.36; t-stat = 7.9). The size of this correlation suggests that an increase in residual temperature value by 20 degrees results in a 0.36 percentage point increase in the convertible percentage of total vehicles sold (a 14.4% increase given the baseline of 2.5%).[8]

[8] We use 20 degrees as a convenient way to think about the overall size of the effect. The extremes of the data are temperature residuals of approximately -20 and 20 degrees. Thus, 20 degrees can be thought of as an extreme temperature value in the data relative to average, or the difference between having a somewhat lower temperature value than average (-10 degree residual) compared to a somewhat higher temperature value than average (10 degree residual).

A natural question is whether abnormally high temperature is only effective in the early spring. In other words, people may buy a convertible as soon as it warms up in the spring time—but may not be impacted by abnormal temperature variation in the fall. Figure 5 provides the scatter plots for abnormal temperature and abnormal convertible sales in Chicago separately for each quarter of the year. The results suggest a strong and significant positive correlation in quarters 1, 2, and 4 (t-stats: 5.2, 4.0, and 4.7 respectively). We argue that the lack of statistically significant correlation in quarter 3 (July, August, and September) likely reflects the fact that since the weather is already so warm during quarter 3, abnormally high temperature does not increase the instantaneous utility for owning a convertible—a necessary condition for projection bias to cause an increase in purchases. Particularly important, however, is the strong positive and significant correlation in quarter 4. Similar to springtime, a week with abnormally warm weather during November in Chicago results in a large increase in the percentage of convertibles sold.

The impact of abnormal weather variation on convertible sales that we find in Chicago may not generalize to all types of DMAs. We use Miami as the second example for how weather impacts convertible sales. Figures 6, 7, and 8 replicate Figures 3, 4, and 5 using data from the Miami-Ft. Lauderdale DMA. Figure 6 illustrates a much weaker seasonal pattern in convertible sales in Miami than was found in Chicago. In addition, Figure 7 shows that the mean daily temperature is both warmer on average and less variant in Miami than in Chicago. Relatedly, the deviations of weather in Miami from weekly norms (Panel B of Figure 7) are smaller than in Chicago (Panel B of Figure 4). Due to the warmer average temperature in Miami, we would predict that abnormally warm weather in Miami does not increase the fraction of vehicles purchased that are convertibles by nearly as much as abnormally warm weather in Chicago. To test this directly, we calculate the correlation between the residual convertible sales in Panel B of Figure 6 and the residual temperature from Panel B of Figure 7. The overall correlation between residual convertible sales and residual temperature in Miami is actually negative (although not statistically significant; t-stat: -0.5). Figure 8 shows that the correlation is not statistically significant for any of the four quarters of the year.

We generalize from our Chicago versus Miami example by combining the data for all DMAs to estimate the impact of temperature on convertible sales across our entire sample. We do so by estimating the following specification.

$$PercentConvertible_{rt} = \alpha_0 + \alpha_1 Weather_{rt} + \mu_{rt} + \tau_{rT} + \epsilon_{rt} \quad (4)$$

PercentConvertible measures the percentage of vehicles sold in DMA *r* during week *t* that were convertibles. *Weather* is a vector of weather variables for DMA *r* in week *t*—temperature, rainfall, snowfall, slushfall and cloud cover—defined previously in this section. (Summary statistics can be found in Table 9.) μ_{rt} are DMA*week-of-the-year fixed effects and τ_{rT} are DMA*year fixed effects. Given the varied size of the DMAs in our sample, we weight the regression based on the total number of vehicles sold in the DMA-week.

Table 1 reports the results of estimating Equation (4). Column 1 indicates that when the temperature is 1 degree higher than expected in a given DMA, the DMA experiences on average an increase of 0.011 percentage point in the convertible fraction of total vehicles sold. Thus a 20-degree swing in temperature in any given week, is predicted to change the convertible percentage of vehicles sold by 0.22 percentage points (an 8.5% change relative to the weighted base rate of 2.6% of vehicles sold being convertibles). Liquid inches of rain, snow, and slushfall all have negative impacts on the convertible percentage of vehicles sold, although these effects are relatively small given the amount of variation in rain, snow, and slushfall that exists in the data. Cloud cover is also very important for convertible demand. As the sky goes from completely clear to completely cloudy, convertible sales decrease by 0.172 percentage points. Thus, a clear sky (relative to completely overcast) increases convertible demand by the same amount as approximately 16 degrees higher temperature.

Another way to understand the size of the estimated effect would be to calculate the decrease in price that would be necessary in order to reduce sales by the same amount. Berry, Levinsohn, and Pakes (1995) estimated demand elasticities for 13 specific models of vehicles. Their estimates ranged from approximately -3 to -6. Assuming an average convertible price of $40,000 (the average in our data), price would have to fall by $1,133 (assuming an elasticity of demand of -3) to $567 (assuming

an elasticity of demand of -6) in order for quantity demanded to fall by 8.5 percent, the predicted effect of a 20-degree change in temperature. This suggests that the size of the effects we find are much larger than the utility most people would get from owning a convertible during one week of particularly good weather.[9]

The next four columns in Table 1 break down the impact of temperature and other weather variables on convertible sales by quarter of the year. Consistent with the Chicago example shown in Figures 3 through 5, the effect of temperature is large and statistically significant in quarters 1, 2, and 4, but insignificant in quarter 3 (when baseline temperature is already quite warm in most areas). Cloud cover—which is arguably important no matter what time of year—is large and significant in all quarters (including quarter 3).

As our Chicago and Miami examples illustrate, the overall effects that we present in Table 1 are likely to mask important heterogeneity that exists in the data. To better understand this heterogeneity, we estimate the impact of temperature on convertible sales separately for DMA-weeks-of-the-year with different mean values for the daily high temperature. The mean value for the daily high temperature for each DMA-week-of-the-year was obtained by calculating the average of the daily high temperatures in a given DMA-week-of-the-year across the different years in our sample. We then group DMA-weeks into 5-degree bins by average daily high temperature for the corresponding DMA-week-of-the-year. We re-estimate Equation (4) for each bin. Figure 9 plots the temperature coefficients (with 95% confidence intervals) estimated for each 5-degree bin of average temperature values. For example, the leftmost point plotted in the graph is the estimated coefficient for DMA-weeks-of-the-year whose average daily high temperature is less than 35 degrees. This figure illustrates that abnormally high and low temperature values have large and significant impacts on convertible sales when the baseline temperature for a given DMA-week-of-the-year is less than 80-85 degrees. The point estimates for these degree bins range from 0.010% to 0.019%. As the average daily high temperature rises above 80 degrees, however, we find that abnormal temperature

[9] This is particularly true if one considers the possibility of renting a convertible in order to enjoy a week of unusually good weather, an option that would be less hassle and would not require buyers to bear the initial depreciation associated with buying a new vehicle.

variations have little effect on convertible sales. In fact, we find negative values at the very highest temperature ranges suggesting that an increase (decrease) in mean daily temperature over these hot baselines may have a negative (positive) impact on convertible purchases. These heterogeneous effects explain the zero-effect of temperature on convertible sales that we found for Miami since Miami's expected temperature is nearly always above 80 degrees.

Baseline 4-Wheel Drive Results. While buying a convertible may seem especially attractive on a warm day, it is cold and snowy days that make 4-wheel drive vehicles seem like an especially good idea. Table 2 presents our estimate of the impact of weather variation on the 4-wheel drive percentage of total vehicles sold obtained by substituting *Percent4WheelDrive*, the percentage of total vehicles sold that are 4-wheel drive, on the left hand side of Equation (4). As we expected, the results we find are the opposite of what we found for convertible sales. We find that colder temperature values lead to more 4-wheel drive purchases. For example a 20-degree change in temperature leads to a 1.0 percentage point change in the percentage of 4-wheel drive vehicles purchased (a 3% change relative to the weighted baseline of 33.5% of vehicles sold with 4-wheel drive). We also find a large, positive impact of snow and slush on 4-wheel drive transactions. One inch of liquidized snow (about 10 inches of snow) leads to a 1.02 percentage point increase in the percentage of total vehicles sold with 4-wheel drive. The effects for snowfall are statistically significant in quarters 1 and 4 (the standard errors for quarters 2 and 3 indicate that we do not have sufficient snowfall variation to estimate effects in these quarters). The effect of snowfall is larger in quarter 4 than in quarter 1. However, the significant impact of snowfall in quarter 1 suggests that even a snow storm that occurs towards the end of the winter season can have a powerful impact on 4-wheel drive purchase behavior.

The amount of snowfall each week has a very different distribution from the distribution of temperature. Snowfall is usually zero in most DMA-weeks, but can have very large values in a few DMA-weeks. The nature of this variable suggests a modeling approach along the lines of an event-study design. What happens in the weeks leading up to and after a big snow storm? We present the results from an event-study design in Figure 10. We choose the events to be the largest snow storm

(measured in amount of snowfall) that occurs in each July-to-June year in each DMA in our sample that has above median weather variation. (This excludes places with no snowfall.) We regress the 4-wheel drive percentage of total vehicles sold in each DMA-week on DMA*year and DMA*week-of-the-year, weighting by the total number of vehicles sold in each DMA-week. We obtain the residuals from this regression for each observation, and sort the residuals by the number of weeks before or after the largest snow storm of the year in the DMA where the observation occurred. Figure 10 plots the average of these residuals for the 12 weeks before and the 12 weeks after each of these events. As can be seen in Figure 10, we find limited evidence that individuals increase their 4-wheel drive purchases leading up to a snow storm. We then see a large spike at the event date such that the percentage of vehicles sold that have 4-wheel drive goes up by almost 1 percentage point. This effect diminishes but continues to be significant for two more weeks before returning to baseline.

Our analysis uses the percentage of total vehicles sold with 4-wheel drive as the outcome of interest. Thus, a change in this measure can be due to an increase in 4-wheel drive purchases or a decrease in purchases of vehicles without 4-wheel drive. Analysis on the log number of convertible and 4-wheel drive purchases made confirm the finding that convertible purchases increase substantially during warm-weather weeks, but show that 4-wheel drive purchases actually decrease during and after snow storms—but not by as much as purchases of vehicles without 4-wheel drive. Thus, it is worth noting that the 4-wheel drive results are driven in part by a drop in overall volume. After a snow storm, an individual who is going to purchase a 4-wheel drive vehicle appears to be more motivated go to the dealership than buyers of non-4-wheel drive vehicles.

Alternative Characteristics. We have estimated the effect of weather on the sales of vehicles with two characteristics whose utility is weather-related, convertible roofs and 4-wheel drive. One could imagine a variety of other characteristics whose value to a customer also varies with weather: air conditioning, sunroofs, snow tires, towing packages, etc. We cannot estimate the weather-related effects of all of these characteristics because some do not vary much in the data (air conditioning) and others we don't observe (sunroofs).

We briefly present the effect of one additional characteristic that we can observe and that varies in the data.[10] Light colors reflect solar radiation, while dark colors absorb it. This means that a black car can be oppressively hot and stuffy if it has been parked outside on a hot and sunny day. Car buyers seem to be familiar with this. Overall in our data, 12.6% of vehicles sold are black; however in Las Vegas, only 9.3% of vehicles sold are black, while in Phoenix the percentage is only 7.8. In Table 3, we report the results of regressing the percentage of vehicles purchased in a DMA-week that are black in color on weather variables, and on DMA*year and DMA*week-of-the-year fixed effects (the same specification as in Tables 1 and 2).

We find, in column 1, that the fraction of vehicles purchased that are black decreases by 0.013 percentage points for every degree increase in temperature. This means that a 20-degree increase in temperature would be associated with a 0.26 percentage point decrease in the sales of black vehicles, a 2.1% change relative to the baseline percentage of 12.6%. Sunshine matters, too. Going from an overcast week to a completely clear week lowers the percentage of black vehicles sold by 0.71 percentage points, or 5.6% relative to the baseline.

In columns 2 through 5 of Table 3, we split the estimates up by quarter. Hot weather and sunny weather reduce the sales of black vehicles in quarters 1, 2, and 4. In quarter 3, we find that sunshine matters even more than in other quarters, while temperature is estimated to matter less.[11]

Dynamic Analysis. The effects that we find for convertibles and 4-wheel drive vehicles suggest that, due to projection bias, idiosyncratic weather differences from week to week can have a large impact on the types of vehicles that people choose to purchase. One concern with this story, however, is that abnormal weather may *appear* to be increasing the demand for certain types of vehicles, but is actually just causing short-run intertemporal substitutions in vehicle purchasing behavior. An example of this "harvesting" story is that a consumer may be interested in purchasing a convertible sometime in the next month and then actually makes her purchase whenever it happens

[10] We thank Loren Pope for suggesting this approach.
[11] In unreported results, we find evidence that cloud cover and temperature are strongly negatively correlated in quarter 3. Specifically, if we estimate the quarter 3 results without cloud cover, the estimated coefficient for temperature increases in magnitude to -0.023 and becomes statistically significant.

to be a nice day outside.[12] In fact, our previously noted finding that abnormally warm weather in November can affect convertible purchases and a snow storm in February can affect 4-wheel drive purchases casts doubt on harvesting as the sole cause of our results. However, these end-of-season purchases cannot rule out harvesting entirely as a contributing factor to our results.

In order to directly address short-run intertemporal substitution of purchases, we estimate a distributive-lag model that adds to the weather variables during the week of purchase a one-week lead and 12 weeks of lagged weather variables. By including lag variables, we are able to test whether having cold or hot weeks leading up to the week of purchase influences how the current weather affects behavior. For example, in the convertible scenario, negative coefficients on the lag variables are interpreted as evidence of harvesting via the following argument. A negative coefficient on, say, the three week lag of temperature indicates that if the weather *three weeks ago* was hot, sales *this week* are lower by some amount than they otherwise would have been. This implies that if the weather *this week* is hot, sales *three weeks from now* will be lower by that same amount. We can thus use the lag coefficients to answer the question "If the weather is hot this week, how much lower will sales be in subsequent weeks?" The one week lag gives us an estimate for the effect of hot weather this week on sales one week from now, the two week lag estimates the effect of hot weather this week on sales two weeks from now, and so on. Thus, if we add up all our lag coefficients and find that they equal the negative of the current period coefficient, it suggests that any increased sales that occur due to hot weather this week are made up entirely of sales displaced from the twelve following weeks. More generally, the sum of the lag coefficients tells us how much of our estimated current period effect is due to intertemporal substitution.[13]

Table 4 presents the results of this dynamic analysis for convertible purchases. The results once again show a large and significant effect of current weather on convertible purchases. The coefficients on the lag variables are all small relative to the current temperature coefficient, in most

[12] The fact that more convertibles are bought in spring than winter and the reverse for 4-wheel drive vehicles suggests that there may be harvesting in response to the overall seasonal pattern of the weather. However, this does not mean that harvesting happens in response to idiosyncratic weather variation.

[13] See Jacob, Lefgren, and Moretti (2007) for a similar analysis that tests for intertemporal substitution of crime using abnormal weather shocks and Deschenes and Moretti (2009) who test for intertemporal substitution of mortality using abnormal weather shocks.

cases not statistically significant, and more often positive than negative. In the full data (Column 1 of Table 4), there is no evidence that warmer than usual weather in the previous weeks affects the current week's sales. If anything, it appears that several weeks of warm weather in a row might lead to an even larger demand for convertibles. There is also no evidence that warm weather in the following week (the lead 1 variable) has a significant impact on current convertible sales, which serves as a nice placebo test.

Table 5 provides a similar analysis for 4-wheel drive purchases. This analysis indicates that snowfall anytime in the last three weeks leads to an increase in the percentage of vehicles sold with 4-wheel drive. There is, however, evidence of some short-run substitution in demand. The summation of the coefficients for lag 4 through lag 12 is -1.13 percentage points. Thus, approximately 47% of the positive effect of snowfall on 4-wheel drive purchases that occurred in the current, lead, and 3 lag weeks (2.42 percentage points) could be considered as arising from harvesting. In other words, the increase in the percentage of 4-wheel drive vehicles purchased after a snow storm is smaller if there was a snow storm that occurred sometime in the previous two to three months. (Presumably, this is because some people purchased a 4-wheel drive vehicle in the wake of the earlier snow storm and no longer need to buy one.) Overall, this dynamic analysis suggests that the increase in demand for convertibles and much of the increase in demand for 4-wheel drive vehicles that we find due to abnormal weather variation cannot be explained by short-run intertemporal substitution in demand.

Test Drive Timing. One aspect of vehicle purchasing that may lead to a correlation between weather and vehicle purchase timing, particularly for convertibles, is the desire of most customers to test drive a vehicle before buying. Suppose a customer is considering buying a convertible, and that she does not suffer from projection bias, meaning that she has no problem accurately forecasting her utility from owning a convertible in various weather states. Now suppose that, before she buys the convertible, she would like to be able to test out various features of the convertible: how convenient it is to put the top up and down, how much wind or road noise she experiences with the top down, etc. It is unpleasant to do such a test drive when the weather is cold, so she waits for a

warm day to go to the dealership, test drive, and ultimately purchase the convertible. Alternatively, suppose that another customer suddenly needs a replacement vehicle, perhaps because his current vehicle has broken down and is no longer worth repairing. Suppose that a convertible is one of the vehicles he would consider purchasing, but on the day he needs the new vehicle it is too cold to test drive a convertible. Unwilling to buy the convertible without being able to test out the convertible features of the car, he buys a non-convertible instead.

The behavior of both of these types of customers would lead to a higher percentage of vehicles sold on warm days being convertibles relative to cold days for reasons other than projection bias. The first type of customer that we outlined above would lead to harvesting (customers wait until a warm week to buy a convertible—so that they can test drive the vehicle). We already discussed and ruled out harvesting effects for convertibles in the previous section. However, the second customer type that we discuss above is not ruled out by our distributive lag model. Several pieces of evidence, however, argue against a test-drive learning story. For example, Figure 9 indicates that an extra degree of warm weather results in more convertible purchases even when the baseline temperature is in the 60-80 degree range. This is a range of temperature for which it is clearly possible for someone to test out the various car features comfortably. Our results thus suggest that it is more than simply testing the features of a car that cause warm weather to result in a higher number of convertibles being sold. We can also get a sense of how important test drive timing might be for our results by considering the effect of cloud cover. There is no reason that a customer could not test drive a convertible on a day that is cloudy—as long as it is not cold or rainy. Thus, in our regressions, which control for temperature and rain, we should not see an effect of cloud cover if the reason for the correlation between temperature and convertible purchases is test drives. However, projection bias should lead to warm, sunny days being days on which people are particularly likely to buy convertibles, rather than warm, cloudy days. Indeed, if we examine the results in Table 1, we find that unusually cloudy days have a significant negative effect on the percentage of vehicles sold that are convertibles, consistent with projection bias. It is particularly noteworthy that cloudy days have a negative effect in all four quarters, and the effect of cloudy days is largest in the third quarter, when

days are generally warm. This third quarter effect is especially suggestive of the fact that people buy more convertibles on warm days not because it is more possible to test drive them, but because it seems more attractive to own a convertible on such days.

Vehicle Buyers who Previously Owned a Convertible or 4-Wheel Drive Vehicle. Another alternative hypothesis that would explain our findings is that customers need to test drive a vehicle on somewhat extreme weather days (warm, sunny ones or cold, snowy ones) in order to actually learn what their utility will be from owning either a convertible or a 4-wheel drive in such weather conditions. Under this hypothesis, a warm, sunny day does not lead a customer to overestimate the utility she will get from owning a convertible; instead it enables her to learn for the first time how high her true utility will be from owning a convertible in such weather states. Before considering this as an alternative hypothesis, we note that this type of extreme learning story—in which vehicle buyers can't quite imagine what it would be like to own this vehicle in another state of the world even when they have experienced that state of the world many times—starts to mesh together with exactly what projection bias is; namely, the inability to appreciate the utility that one will experience when the state of the world changes.

Even though projection bias and learning might look similar at their extremes, our data allow us to investigate somewhat more direct evidence for learning as an explanation. In our data, we observe what trade-in, if any, customers bring when they buy a vehicle. This means we can observe vehicle transactions by customers whom we know have already owned a convertible or have already owned a 4-wheel drive vehicle. Previous convertible owners are less likely to need to "learn" about what it is like to own a convertible during a warm weather state, and similarly for previous 4-wheel drive owners and cold or snowy states, so evidence that abnormal weather impacts these buyers is particularly strong evidence for projection bias.

If we look within the subset of transactions that use a convertible as a trade-in, we find that approximately 25% of these buyers purchase another convertible while 75% purchase a non-convertible vehicle. Column 1 of Table 6 reports the results of our baseline specification if we restrict the sample to buyers who are trading in a convertible. While the standard errors are much

larger due to the sample restriction, we continue to find a positive impact of temperature at the time of purchase on convertible demand. The point estimate is about six times larger than the point estimate in the entire sample—although the larger estimate in percentage *point* terms is smaller in percentage terms because the convertible purchase rate in this sample (25%) is so much higher.[14] In Column 2 of Table 6, we estimate the effect of weather on buyers who are trading in a 4-wheel drive vehicle. Overall, 78% of people who trade in a 4-wheel drive vehicle purchase another 4-wheel drive vehicle. In Column 2 we continue to find strong and statistically significant effects of abnormal weather—including temperature, snowfall, slushfall, and cloud cover—on 4-wheel drive purchases for buyers who traded in a 4-wheel drive vehicle. The estimated effects are substantially smaller in percentage terms than in the full sample, in large part because the unconditional probability of buying a 4-wheel drive vehicle is so high in this sample.

The fact that we find effects of abnormal weather in precisely the subsample of buyers who would seem to have the *least* to learn about their utility from owning either a convertible or a 4-wheel drive vehicle cast doubt on a learning story explaining the effects that we find.

Expensive Vehicles. One might worry that our finding that warm weather leads to higher convertible sales is simply spurious correlation of the following sort. Suppose that good weather puts people in a generally good mood, and that when people are in a good mood they spend money more freely. If that were so, then we might see good weather associated with higher convertible sales simply because convertibles are more expensive on average. (In our data, the average price of a vehicle that is not a convertible is $20,542, while the average price of a convertible is $30,845.)

We investigate this hypothesis by re-estimating Equation (4) (the specification reported in Column 1 of Table 1) replacing the dependent variable "percentage of vehicles sold that are

[14] The full sample results indicate that a 20-degree increase in abnormal temperature increases the percentage of vehicles sold that are convertibles by 0.22 percentage points in the full sample, an 8.5% increase relative to a base percentage of 2.6%. In the "convertible trade-in" subsample, the effect is a 1.2 percentage point increase, a 4.8% increase relative to a base percentage of 25%.

convertibles" with "percentage of vehicles sold whose price is greater than *X*." We estimate four variants of this alternative specification, with *X* equal to $20,000; $30,000; $40,000; and $50,000.[15]

Our original specification found that when the temperature rises by one degree (all else equal), the percentage of vehicles sold that are convertibles goes up by 0.011 percentage points (t-stat = 14.4). In our alternative specification, we find that when temperature goes up by one degree, the percentage of vehicles sold whose price is more than $20,000 is unchanged (coefficient estimate is 0.000, t-stat = 0.28). The estimated coefficient for a threshold price of $30,000 is 0.002 (t-stat = 0.93); for a threshold price of $40,000 the estimated coefficient is 0.001 (t-stat = 1.5); and for a threshold price of $50,000 the estimated coefficient is 0.000 (t-stat = 0.39). These results give no evidence that our original effect is driven by buyers buying more expensive vehicles in good weather.

Returning Vehicles. Projection bias suggests that people can make mistakes when purchasing a durable good and that people may realize the mistake when the state of the world changes. Conlin, O'Donoghue, and Vogelsang (2007) make this case and specifically test for mistakes by analyzing whether cold-weather items (boots, gloves, etc.) purchased by mail order were more likely to be returned if the purchase was made during a very cold state. In the car market, projection bias mistakes might be identified by seeing vehicles that were purchased during abnormal weather weeks reappear in the market (either as trade-ins or as subsequent used car sales) more quickly than vehicles that were purchased during normal weather weeks. The quick return of a vehicle to the market could indicate that the owner was not happy with the purchase he or she made.

Unfortunately, there are at least two reasons why testing for early returns in the vehicle market is much harder than for catalog orders. First, is simply a data limitation. Although our data are impressive and represent a 20% sample of all new car dealerships in the U.S., we can only identify "returned" vehicles that happen to be traded in or sold as a used vehicle at one of the dealerships we observe. Said another way, for any vehicle whose sale we observe at some point, we have roughly a 20% chance of seeing that vehicle's subsequent return or resale if that transaction happens at a

[15] We run this analysis on all non-convertible vehicles. We eliminate convertibles from this analysis so that our results are not affected by our finding that higher temperatures are associated with increased convertible sales. For this analysis, we want to know whether higher temperatures increase the sales of high-priced vehicles absent a convertible effect.

dealership, and no chance of seeing it if that transaction happens person-to-person. Second, and perhaps more importantly, car dealerships do not offer the kind of "no-hassle return" policies that are common for catalog retailers. A mistake that is made when buying gloves can be easily fixed with a few minutes and a little postage. However, an individual who realizes that he or she has made a mistake after buying a convertible cannot return it so easily. To switch the convertible for a hardtop will likely require that the individual sell the convertible (likely at a loss if the vehicle is new because of the rapid initial depreciation of new vehicles) and buy the hardtop at full price. Thus, even if mistakes are being made, the mistakes may not be large enough to merit fixing.

Despite these two concerns, we test for the impact of abnormal weather at the time of purchase on how quickly the vehicle reappears in the market. Of the roughly 40 million vehicles that are transacted in our dataset, 2.37% of them reappear within 1 year as a trade-in or subsequent sale, 5.03% within 2 years, and 7.16% within 3 years.[16] On average in the U.S., owners keep their vehicles for just over 5 years (Polk, 2010).

Our empirical strategy is to estimate whether convertibles that were purchased when the weather was abnormally warm and 4-wheel drive vehicles that were purchased when the weather was abnormally cold are more likely to reappear in our data within a short time frame than vehicles purchased under more typical weather conditions. The columns of Table 7 report results for regressions in which the outcome variable is an indicator that equals one for a given transaction if we observe the transacted vehicle reappear in our data as a trade-in or in another sales transaction within, respectively, 1, 2, or 3 years. We control for DMA*week fixed effects to eliminate seasonal and geographic differences in how quickly vehicles are returned. Table 7 shows that convertibles are, overall, 1.272 percentage points more likely to be returned within a year than other types of vehicles; 4-wheel drive vehicles are also more likely to be returned (by 0.285 percentage points) than other types of vehicles. The positive signs of the coefficients estimated for the interaction of convertible and temperature variables are consistent with projection bias: convertibles are more likely to be returned quickly when they were purchased during abnormally warm weather weeks. However, this

[16] Unique identification numbers corresponding to individual VIN numbers are used to track vehicles over time.

result is statistically significant only in column 2. The point estimates suggest that when the weather is 20 degrees warmer, convertibles are 0.34 percentage points more likely to be returned within 2 years than hardtops (a 4.6% change relative to the baseline convertible return rate of 7.332%). The temperature interaction with 4-wheel drive vehicles is more consistently statistically significant, and indicates that a 4-wheel drive vehicle is more likely to be returned within 1, 2, or 3 years if it is purchased in an abnormally cold week. Overall, our results for the effect of abnormal weather on returning vehicles, while clearly suggestive, is less strong than our evidence for the effect on purchasing vehicles. An important contributor to this is simply that the number of vehicles we see sold and then see reappear within our data is not that high. As a consequence, we have limited ability to identify differences in the rates at which vehicles are returned under different circumstances.

Price Effects. We have shown in the previous sections that the percentage of vehicles sold that are convertibles is higher in weeks with warm and sunny weather, while the percentage of vehicles sold that are 4-wheel drives are higher during and just after weeks with cold, snowy weather. We argued that this is evidence of projection bias—that individuals are over-influenced by the current weather when they are making vehicle purchase decisions.

At a market level, one could describe this effect as an increase in demand associated with unusually warm and sunny, or unusually cold and snowy, weather. Thinking of the phenomenon this way, one might wonder whether there is an effect of projection bias not only on the quantities of vehicles people buy, but on the prices they pay. In a simple demand model, if the demand curve shifts out while an (upward-sloping) supply curve stays fixed, one would expect to see both higher prices and higher sales quantities.

There are several ways in which this simple model is not an ideal fit for the car industry. First, from a dealer's perspective, the supply of vehicles is not upward-sloping. Dealers can order vehicles from manufacturers as a fixed per unit invoice price in whatever quantity they wish. This corresponds to horizontal marginal cost curve for the dealer. If the dealer is selling vehicles in

competitive market, the effect of an increase in demand should be increased sales, with essentially no increase in price.[17]

Second, a competitive price-taking market is not a very good description of the retail car industry. Individual buyers negotiate a price for a specific vehicle with the dealer. Whether the incremental buyers who are buying as a consequence of projection bias pay higher prices or lower prices than other buyers depends on the reservation prices and bargaining characteristics of projection bias buyers relative to other buyers. One might argue that projection bias buyers must have higher reservations prices than buyers on average, because they are being strongly swayed by temporary weather conditions. Similarly, one might argue that buyers who can buy "on impulse" must have high liquidity, and therefore likely higher incomes and higher reservation prices, than average buyers. Alternatively, one might argue that projection bias buyers are buyers who would not be buying this vehicle on another day, and that the influence of the weather has nudged them just above their point of indifference about buying. In this case, they might well have lower reservation prices than average buyers. Similarly, if dealers recognize which buyers are projection bias buyers, they may realize that they must offer a good price today, or lose the sale forever, since in another few days the weather will change and these buyers will no longer be in the market.[18]

Overall, we conclude that it is an empirical question whether prices for convertibles and 4-wheel drives will be higher in the same weeks that warm and sunny weather or cold and snowy weather leads to increased sales of these types of vehicles. We estimate the effect of weather on the prices of convertibles and 4-wheel drives using the following specification.

$$Price_{ijrt} = \beta_0 + \beta_1 Weather_{rt} + \beta_2 PurchaseTiming_{it} + f(Odometer_i, \beta_3) + \mu_{rt} + \tau_{rT} + \phi_i + \epsilon_{ijrt} \quad (5)$$

Price measures the price paid in transaction *i* for vehicle *j* that occurred during week *t* in DMA *r*. (In order to make our measure of price represent a customer's total wealth outlay for the vehicle, we

[17] Dealers place orders for vehicles months in advance, so over a horizon of several months, a dealer's supply of vehicles is predetermined. However, dealers can sell more or fewer vehicles on any given day, meaning daily vehicle supply is not fixed. (For more on how dealer supply and inventory affects prices, see Zettelmeyer, Scott Morton, and Silva-Risso (2007).

[18] We thank Glenn Ellison for suggesting this point.

define price as the contract price for the vehicle agreed upon by the buyer and the dealer, minus any manufacturer rebate the buyer received, plus any loss (minus any gain) the buyer received in negotiating a price for his or her trade-in.) *Weather* is a vector containing the temperature, rainfall, snowfall, slushfall, and cloudcover for week *t* in DMA *r*. *PurchaseTiming* is a vector containing indicators for whether transaction *i* occurred during the weekend, or at the end of the month, times in which salespeople may be willing to sell vehicles at a discount in order to hit sales volume targets. The specification also includes DMA*year (τ_{rT}), DMA*week-of-year (μ_{rt}), and "vehicle type" (ϕ_i) fixed effects. (A vehicle type is defined by the interaction of make, model, model year, trim level, doors, body type, displacement, cylinders, and transmission.) We estimate Equation (5) separately for new convertibles, used convertibles, new 4-wheel drives, and used 4-wheel drives. The specifications that estimate the effect of weather on used vehicle prices also include a linear spline in the vehicle's odometer, which allows vehicle prices to depreciate over time in a reasonably flexible way. (See Busse, Knittel, and Zettelmeyer (forthcoming) for use of a similar specification to estimate price effects in similar data.)

Table 8 reports the results of estimating equation (5). Generally speaking, we find that the effect of weather on prices is fairly small, even when it is statistically significant. In Column 1, which estimates the effect of weather on new convertible prices, none of the weather variables have statistically significant effects. In Column 2, which estimates the effect of weather on used convertible prices, an increase in temperature of 20 degrees would be estimated to increase the average price of convertibles sold during that week by $79.60, a very small amount compared to an average transaction price of $22,222 for used convertibles. In addition, one inch of liquidized snow (about 10 inches of snow on the ground) is associated with transaction prices that are lower by $114.48. For 4-wheel drives, the results are similarly small, and the directions of the effects are mostly counter-intuitive. A 20-degree drop in temperature would be predicted to decrease the average transaction price of a new 4-wheel drive by $16.60 and of a used 4-wheel drive by $40.60. These are very small effects relative to an average transaction price of $31,845 for new 4-wheel drives and $19,132 for used 4-wheel drives. In addition, for used 4-wheel drives, a 10-inch snowfall

is predicted to decrease the average transaction price by $23.80, while going from a sunny to an overcast week is predicted to decrease the average transaction price by $54.06. The only coefficient that is statistically significant in the expected direction is the estimate of the effect of cloud cover on new 4-wheel drive prices; going from a completely sunny week to a completely overcast week is predicted to increase the price of a new 4-wheel drive by $36.44.

III. Housing Market

Data and Empirical Strategy. Our analysis is based on a housing dataset of more than four million single-family residential properties across the United States that sold at least twice between 1998 and 2008. We purchased the data from a commercial vendor who had assembled these data from assessors' offices in individual towns and counties.[19] Since larger metropolitan areas are more likely to archive their assessor data electronically and sell it to commercial vendors, urban counties are over-represented relative to rural counties.[20] The data include the transaction price and the sale date of each house, the previous transaction price and sale date, a physical address (from which we obtain county and state indicators), and a consistent set of structural characteristics, including swimming pool, central air, fireplace, lot size, year built, square feet of living area, number of bathrooms, and number of bedrooms. We observe transactions at two different dates for a single house, but only a single set of characteristics for a given house—the characteristics that existed at the time of the second transaction.[21]

While we observe the closing date for each home, we do not observe the date that the home went under contract—which is the relevant date for testing a model of projection bias. Throughout the analysis we assume that homes go under contract two months prior to the closing date. We base this assumption off of a small dataset consisting of homes in the Chicago area for which both

[19] The commercial vendor is Dataquick which is the source of housing data for many papers in the literature.
[20] Certain states are overrepresented in the data. For example, 30.7% of sales were in California, 14.1% in Florida, 8.9% in Ohio, 6.9% in Washington, 4.8% in Massachusetts, and 4.2% in Nevada. Data also include observations from AL, CO, CT, GA, HI, IA KY, MI, MN, MO, NC, NE, NY, OK, OR, PA, RI, SC, TN, TX, and VA.
[21] This is because we are essentially creating two transactions from a single observation—an observation which records the current sale price and sale date of a particular house, and the sale price and sale date of the most recent previous transaction for that same house if the previous transaction occurred during our data window.

contract date and closing date were available.[22] In this dataset, the average time between contract and closing dates was 52 days. After including a few days for price negotiations, we assume that purchase decisions were made on average two months prior to the closing date for each house. Because we lack exact data for the day the home went under contract, our empirical strategy in this section of the paper will be restricted to examining the effect of seasonal weather patterns rather than precise idiosyncratic weather differences at the time of purchase.

We clean the data in a similar manner as previous work that has used this type of housing data in order to eliminate outlying observations. Housing transactions are dropped if the sales price was less than $5,000 or more than $5,000,000, if the house was built before 1900, if the square footage is less than 250 square feet or more than 10,000, if the number of bathrooms is less than 0.5 or more than 10, and if the number of bedrooms is less than 1 or more than 10. We also drop all new construction (age less than 2 years old).[23] We also restrict the sample to houses located in counties that report whether a home has a swimming pool. As described in our empirical section below, in order to perform a repeat-sales analysis, our sample contains only transaction for houses that had a previous sale within our sample period.

Table 10 provides some basic summary statistics for the final set of housing transactions in our dataset (roughly 4.2 million observations). The average sales price over the entire time frame of our data was approximately $275,000, again reflecting the fact that urban areas (and California) are overrepresented in our housing data sample relative to the entire population of housing within the United States. The Table also shows that 12% of homes in our sample had swimming pools, 30% had central air, and 46% had at least 1 fireplace.[24] The average home was built in 1968 on a 0.32 acre lot with approximately 3 bedrooms, 2 baths, and 1,700 square feet of living area.

---

[22] We thank Steve Levitt and Chad Syverson for sharing this information with us.

[23] We do not have an indicator in the dataset for when a home is being sold for the first time. One potential concern is that new homes (which sell for a premium) may be more likely to have a swimming pool and may also have a strong seasonal pattern (which could bias in favor of the results we find). Because we lack an indicator for new homes, we simply drop homes that may fall in this category.

[24] Certain counties in our dataset reported no homes as having either central air or a fireplace (which provides an indication that these data were not systematically collected in those counties). Given that all of our analysis has county fixed effects, we leave these homes in the dataset in order to provide more observations for other housing characteristics of interest (e.g. swimming pools).

Our goal is to test for the presence of projection bias using a very simple empirical strategy and to provide the results in a graphical fashion. Our primary specification is

$$Log(Sales\ Price)_{itc} = \gamma_i + \theta_{tc} + \varepsilon_{itc}, \quad (6)$$

where $Log(Sales\ Price)_{itc}$ is the log sales price of house i in sample-month t in county c. γ_i is a fixed effect for house *i,* which we can estimate because we use only houses we observe being sold more than once. θ_{tc} is a county*sample-month fixed effect. The residual from Equation (6), e_{itc}, represents, for each house transaction, how much more or less the house sold for than would have been expected after considering how much that very house sold for on another occasion, and how much other houses in the same county sold for during the same sample-month. We will analyze the residuals from this regression by month-of-the-year and house type to see whether there is evidence of projection bias.

Results. We begin by calculating the average residual (obtained from Equation (6)) for homes with swimming pools by the month-of-the-year the house is assumed to have gone under contract (i.e. two months before the sale date). In Panel A of Figure 11, we plot these average residuals along with their 95% confidence intervals. It is worth noting that these average residuals by month-of-the-year for homes with swimming pools (if weighted by the number of transactions in each month-of-the-year) would sum to zero across months because house fixed effects are included in the regression. Similarly, the residuals for all homes sold within a single sample-month must sum to zero because sample-month fixed effects are included in the regression. Therefore, whenever we see a positive average residual in a given month-of-the-year for houses with swimming pools, we know that the average residual for houses without swimming pools must be negative (although the magnitude of the average negative residual for houses without swimming pools is on the order of one-tenth the size of the average positive residual for houses with swimming pools since houses with swimming pools only represent about 12% of the data).

Panel A of Figure 11 provides the first evidence that swimming pools add more value to a home in the summertime than in the wintertime. Specifically, homes with swimming pools that go under contract in the three hottest months of the year (June, July, and August) sell for 0.22

percentage points more on average than otherwise expected (this effect is jointly statistically significant and individually significant for June and August), while homes with swimming pools that go under contract in the three coldest months of the year (December, January, and February) sell for on average 0.18 percentage points less than otherwise expected (this effect is individually significant for December). Given that the average transaction value of houses with swimming pools in our data is about $398,000, this represents a roughly $1600 swing in value for homes with swimming pools that go under contract in the summertime relative to the wintertime.

Our finding that transaction prices are higher for houses with swimming pools that went under contract in the summer (especially in August) argues strongly against standard discounting or present-biased preferences as the reason for our results. The houses that we identify as selling in August are houses that will close in October, meaning that the buyers of those houses will move in just at the point in the year in which swimming pool season is the farthest away.

One concern with this simple analysis is that while it is clear that the residuals for homes with swimming pools are showing a seasonal trend, it may not be the swimming pool that is causing the seasonal trend, but rather something else about homes with swimming pools. For example, perhaps the seasonal differences are being driven by large homes, which may be more likely than small homes to have swimming pools. To assuage this concern, we regress the residuals from Equation (6), e_{itc}, on all the house characteristics we observe, plus interactions for months-of-the-year. To be precise, we estimate:

$$(7) \quad e_{itc} = \theta_0 + \theta_{1,t}\,\mu_t\,SwimmingPool_i + \theta_{2,t}\,\mu_t\,CentralAir_i + \theta_{3,t}\,\mu_t\,Fireplace_i + \theta_{4,t}\,\mu_t\,LotSize_i + \theta_{5,t}\,\mu_t\,Bedrooms_i + \theta_{6,t}\,\mu_t\,Bathrooms_i + \theta_{7,t}\,\mu_t\,SquareFootage_i + v_{itc}$$

e_{itc} is the estimated residual from Equation (6), and represents how much the log price observed for the sale of house *i* in county *c* in month *t* differs from what would be predicted from other sales of that house and from the sales of other houses in the same county and month. *SwimmingPool, CentralAir,* and *Fireplace* are indicator variables recording whether house *i* has the corresponding feature. *LotSize* measures the size of the lot in acres. *Bedrooms* and *Bathrooms* count the number of

rooms of each type. *SquareFootage* measures the size of the house in square feet. μ_t is an indicator for the month-of-the-year in which the transaction occurs. The coefficients can be interpreted as follows: $\theta_{1,1}$ estimates how large on average the residual of log price (net of house and county*sample-month fixed effects) is for houses with swimming pools that sell in January, conditional on all the other house attributes we observe.

Panel B of Figure 11 presents the twelve swimming pool coefficients ($\theta_{1,1}$ through $\theta_{1,12}$) from Equation (7). Controlling for the seasonal effect of the other housing characteristics on the residual log price does not substantially change our estimates of the effect of swimming pools. Our results continue to show that the value of a swimming pool is higher in the summertime than in the wintertime, although with somewhat reduced statistical significance.

A common procedure when running hedonic models involves trimming the data to eliminate extreme residual values. For example, if the data suggest that a house sold for $100,000 and then sold two years later for $800,000, it is reasonable to assume that there was a data mistake or that the house was changed in a major way. To remove these types of observations, we trim the data to eliminate the top and bottom 1% of residual values and the top and bottom 5% of residual values. Because sales price in Equation (6) is measured as log price, the residual values are also measured in logs. Removing the top and bottom 1% of residual values eliminates homes whose sales price was about 60% more than or 60% less than what would be predicted Equation (6). Removing the top and bottom 5% of residual values eliminates homes that sold for about 25% more or less than Equation (6) would predict.[25]

Figure 12 displays the twelve swimming pool coefficients ($\theta_{1,1}$ through $\theta_{1,12}$) obtained by estimating Equation (7) for the 1% trim sample (Panel A) and the 5% trim sample (Panel B). The same general seasonal pattern for the value of swimming pools remains when trimming the data in this manner. The major advantage to this trimming is that the confidence intervals become much tighter due to the elimination of these high-variance observations. Our preferred specification (with

[25] The 5% and 1% cutoffs for trimming are symmetric because our data consists of exactly 2 observations for each house and we include house fixed effects in our regression. Therefore, every observation in our sample with a positive residual has an observation in the data with a residual of the same magnitude but of the opposite sign.

the 5% trim), provides precise month-to-month point estimates for the value of a swimming pool and shows consistently higher values for homes with swimming pools that sold in the summertime (especially August) when compared to those same homes that sold throughout the wintertime (November through March).

Along with swimming pools, we observe three additional housing characteristics in our data that we believe could have a seasonal component: central air conditioning, fireplaces, and lot size. In Figure 13, we report the estimated coefficients from Equation (7) associated with each of these characteristics, estimated on the 5% trim sample. Panel A shows the estimated coefficients for central air ($\theta_{2,1}$ through $\theta_{2,12}$). There appears to be a seasonal pattern in which central air is worth more in the summertime (especially June and September) and less in the wintertime (November and January-March). The results are smaller in magnitude than those found for swimming pools. Panel B and Panel C present the results for fireplaces and lot size, respectively. The results are smaller in magnitude and less statistically significant than the central air results in panel A. There is little evidence of a discernible pattern in either of these results.

Why do we find small or no results for fireplaces and lot size? It could be that the instantaneous consumption value that these other characteristics provide to homeowners does not vary with season as much as the consumption value of swimming pools across seasons. People may enjoy using fireplaces from fall straight through to spring, and the value of having a large lot may be high both in the spring or fall when yards are very beautiful, and in the summer when people spend a lot of time outdoors.

In Figure 14, we report the seasonal value of other housing characteristics in our data which are unlikely to have a strong seasonal component (number of bedrooms, number of bathrooms, and square footage). We find little evidence of a statistically or economically significant seasonal pattern for these housing characteristics. The lack of seasonal variation in the value of these characteristics (both in terms of statistical significance and effect size) lends credibility to the effects that we find for swimming pools and central air.

Our hypothesis is that higher temperature levels at the time of the purchase decision lead to higher sales prices for houses with swimming pools and central air when compared to purchase decisions made during colder parts of the year. Up to this point, however, we have not used exact temperature, but rather have been using month-of-the-year as a proxy for temperature. Given the variation in weather that exists across the U.S. and across different years in our sample, month of the year is clearly not a perfect proxy for temperature. We remedy this by merging in weather data for every county*sample-month, which allows us to know the average daily high temperature for the month and location in which each house in our dataset went under contract.[26]

The underlying model for how weather and housing characteristics such as a swimming pool interact to impact housing sales is not obvious. For example, it could be that a swimming pool becomes more valuable for every 1 degree increase in temperature. Alternatively, the value of a swimming pool may be constant until the high temperature reaches some hot tipping point (e.g. 70, 80, or 90 degrees). In light of this, we estimate the following specification, whose results are reported in Table 11.

$$e_{itc} = \delta_0 + \delta_1 \, Temp_{tc} \, X_i + \delta_2 Temp_{tc} + \delta_3 X_i + \xi_{itc} \tag{8}$$

e_{itc} is the residual of the log sales price (net of house and county*sample-month fixed effects) obtained from Equation (6). X_i is a vector of the housing attributes we observe (swimming pool, central air, fireplace, lot size, bedrooms, bathrooms, and square footage). We measure $Temp_{tc}$ in four different ways. In Columns 1 and 2, $Temp_{tc}$ is the average daily high temperature in county *c* in month *t*, the month in which the house is inferred to go under contract. In the next three pairs of columns, $Temp_{tc}$ is an indicator variable that corresponds to whether the average daily high temperature in county *c* in month *t* is at least 70, 80, or 90 degrees, respectively. The first column in each temperature pairing in Table 11 reports results for the full sample while the second column reports results for the 5% trim sample. We multiplied all coefficients in Table 11 by 100 for ease of

[26] The temperature information comes from the PRISM Climate Group based at Oregon State University, which provides consistent weather information all across the United States. More information on the weather data we use can be found at http://www.prism.oregonstate.edu/. We accessed the data on 3/12/2011.

reporting. We can therefore interpret—as we do in the next paragraph—the coefficients as approximate percentage point changes.

The first column in Table 11 indicates that for every 1 degree Fahrenheit increase in the average daily high temperature during the month in which the house went under contract, a swimming pool increases the sales price by 0.013 percentage points. This means that a house that sold when the average daily high temperature was, for example, 80 degrees sold for 0.65 percentage points more than the same house that sold when the average daily high temperature was 30 degrees. This effect is statistically significant and remains large and statistically significant when trimming the data to eliminate the top and bottom 5% of residual values (Column 2). In Column 2, central air is also estimated to be more valuable during high temperature months. The interaction effects of temperature and the remaining housing characteristics in these two columns are nearly all small and statistically insignificant.

The next three pairs of columns in Table 11 show the impact of the average daily high temperature being above a threshold of 70, 80, or 90 degrees. Once again we find large and mostly statistically significant effects for the value of a swimming pool. For example, the final column in the table suggests that houses with swimming pools that went under contract in a month where the average daily high temperature was more than 90 degrees sold for 0.37 percentage points more than when these same houses went under contract in a month whose average temperature did not reach 90 degrees.

Although our housing results suggest that projection bias is at work in this market much as we found in the car market, our analysis would be even more compelling if we could see if houses with swimming pools that went under contract in the summertime were more likely to "fall through" and not actually close. This would be analogous in some ways to our results on returning vehicles and to Conlin, O'Donoghue, and Vogelsang's (2007) results on returning cold weather catalog items. Unfortunately our data preclude us from doing such an analysis since we don't have information on homes that went under contract but then did not close. However, this would be an interesting extension if one were able to acquire the relevant housing information to perform this test.

IV. Conclusion

Many of the most important decisions that we make in life involve predicting our future preferences. This paper provides evidence that projection bias may limit our ability to make these predictions accurately. We show that projection bias causes consumers in the car and housing markets to make decisions that are overly influenced by the weather at the time of the decision. We argue that our results imply that projection bias can have important implications for large-stakes markets and that this psychological bias merits additional study and attention.

From a policy perspective, our results suggest that consumers would benefit from laws designed to help them better evaluate their decisions. For example, laws that allow consumers a "cooling-off period" for durable goods or goods for which consumers sign extended contracts may provide significant benefits to consumers. Such laws could also provide incentives for sellers to help buyers be in a "cool" state before an important transaction or contract is made.[27] The Federal Trade Commission has an explicit "Cooling-Off Rule" that applies to situations when "[you] buy an item in your home or at a location that is not the seller's permanent place of business."[28] This rule was made specifically to deal with high-pressure sale situations such as door-to-door sales. The Federal Trade Commission's cooling-off rule does not apply to real estate and automobile sales even though there clearly can be high-pressure sale situations for these important durable goods. While our results suggest that some consumers might benefit from an opportunity to reverse a decision once they have "cooled-off," applying a cooling-off rule to vehicle purchases would provide other consumers an opportunity to game the system by "buying" a new convertible at the beginning of a holiday weekend and returning it after a few days, claiming to have had a change of heart.

Despite showing that projection bias can impact important consumer markets, there are many questions about projection bias that are left unanswered and that future research may be able to address. For example, it is unclear how easy it is to "de-bias" consumers. It is possible that simply

[27] See Camerer et al. (2003) for an extended discussion about cooling-off periods and their potential applications in settings where people make suboptimal choices.
[28] More information on the Federal Trade Commission's "Cooling-Off Rule" can be found on their webpage at: http://www.ftc.gov/bcp/edu/pubs/consumer/products/pro03.shtm .

providing consumers with information about projection bias or asking them to imagine how they will feel about their purchase in a different state of the world could lead to improved decision making. Another extension of our research that would be particularly useful would be to study projection bias for various other state variables—not just weather. For example, emotional states and states of dependency are likely to influence important decisions like having a baby, whether to get married, and whether to accept a given job offer.

References

Berry, S., J. Levinsohn, and A. Pakes. 1995. "Automobile Prices in Market Equilibrium." *Econometrica*, 63(4): 841-890.

Busse, M., C. Knittel, and F. Zettelmeyer. Forthcoming. "Are Consumers Myopic?: Evidence from New and Used Car Purchases." *American Economic Review*.

Busse, M., D. Simester, and F. Zettelmeyer. 2010. " `The Best Price You'll Ever Get': The 2005 Employee Discount Pricing Promotions in The U.S. Automobile Industry." *Marketing Science,* 29(2): 268-290.

Camerer, C., S. Issacharoff, G. Loewenstein, T. O'Donoghue, and M. Rabin. 2003. "Regulation for Conservatives: Behavioral Economics and the Case for 'Asymmetric Paternalism'" *University of Pennsylvania Law Review*, 151(3): 1211-1254.

Conlin, M., T. O'Donoghue, and T. Vogelsang. 2007. "Projection Bias in Catalog Orders." *American Economic Review*, 97(4): 1217-1249.

Deschenes, O. and E. Moretti. 2009. "Extreme Weather Events, Mortality, and Migration." *Review of Economics and Statistics, 91(4),* 659-681.

DellaVigna, S. 2009. "Psychology and Economics: Evidence from the Field." *Journal of Economic Literature,* 47, 315-372.

Jacob, B., L. Lefgren, and E. Moretti. 2007. "The Dynamics of Criminal Behavior: Evidence from Weather Shocks." *The Journal of Human Resources,* 42(3), 489-527.

Laibson, D. 1997. "Golden Eggs and Hyperbolic Discounting." *Quarterly Journal of Economics, 112(2),* 443-77.

Levitt, S. and C. Syverson. 2008. "Market Distortions When Agents Are Better Informed: The Value of Information in Real Estate Transactions." *The Review of Economics and Statistics,* 90(4), 599-611.

Li, Y., E. Johnson, and L. Zaval. 2011. "Local Warming: Daily Temperature Change Influences Belief in Global Warming." *Psychological Science,* 22(4), 454-459.

Loewenstein, G. 1996. "Out of Control: Visceral Influences on Behavior." *Organizational Behavior and Human Decision Processes,* 65, 272-292.

Loewenstein, G., D. Nagin, and R. Paternoster. 1997. "The Effect of Sexual Arousal on Predictions of Sexual Forcefulness." *Journal of Crime and Delinquency,* 32, 443-473.

Loewenstein, G., T. O'Donoghue, and M. Rabin. 2003. "Projection Bias in Predicting Future Utility." *Quarterly Journal of Economics*, 118: 1209-1248.

Loewenstein, G. and D. Schkade. 1999. "Wouldn't It Be Nice? Predicting Future Feelings." in *Well-Being: The Foundations of Hedonic Psychology*, ed. Daniel Kahneman, Edward Diener, and Norbert Schwarz, 85-105. New York: Russell Sage Foundation.

Nisbett, R. and D. Kanouse. 1968. "Obesity, Hunger, and Supermarket Shopping Behavior." Paper presented at the 76th Annual Convention of the American Psychological Association, San Francisco, CA.

Nordgren, L., J. Van Der Plight, and F. Van Harreveld. 2006. "Visceral Drives in Retrospect: Explanations About the Inaccessible Past." *Psychological Science,* 17, 635-640.

Nordgren, L., J. Van Der Plight, and F. Van Harreveld. 2007. "Evaluating Eve: Visceral States Influence the Evaluation of Impulsive Behavior." *Journal of Personality and Social Psychology,* 93(1), 75-84.

O'Donoghue, T. and M. Rabin. 1999. "Doing It Now or Later." *American Economic Review,* 89(1): 103-24.

Polk. 2010. "Consumers Continuing to Hold onto Vehicles Longer." November 3, 2010. https://www.polk.com/company/news/consumers_continuing_to_hold_onto_vehicles_longer_according_to_polk

Read, D., and B. van Leeuwen. 1998. "Predicting Hunger: The Effects of Appetite and Delay on Choice." *Organizational Behavioral and Human Decision Processes*, 76, 189-2005.

Risen, J. and C. Critcher. 2011. "Visceral Fit: While in a Visceral State, Associated States of the World Seem More Likely." *Journal of Personality and Social Psychology,* 100(5), 777-793.

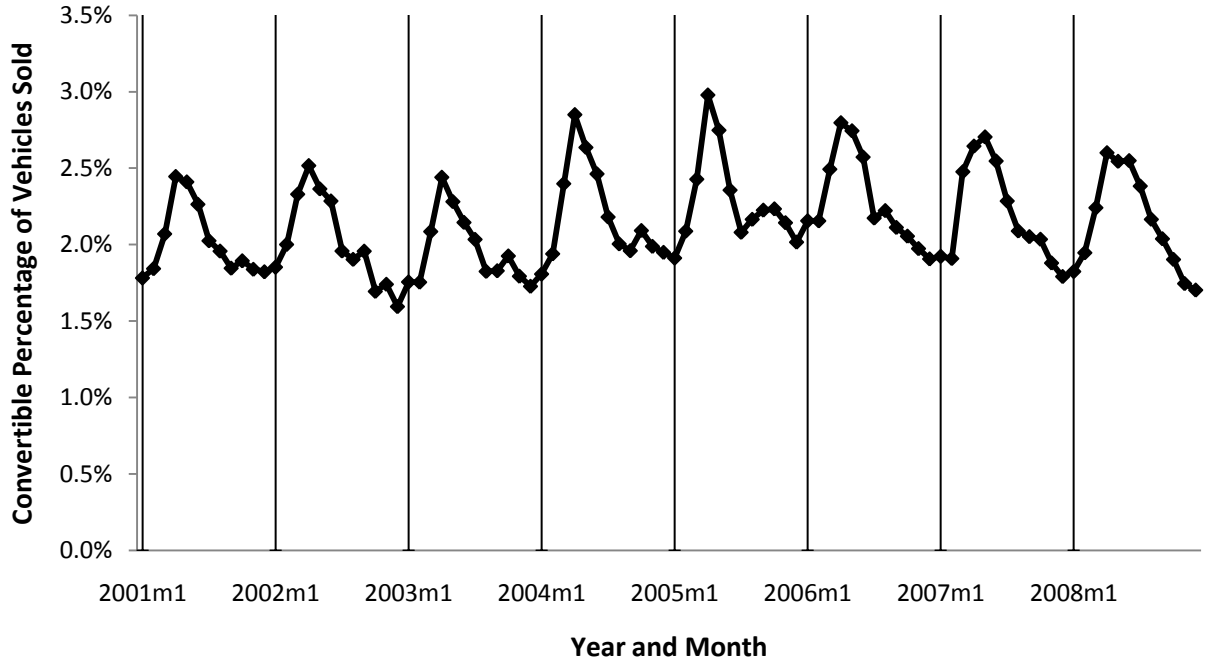
Simonsohn, U. 2010. "Weather to Go to College." *Economic Journal,* 120(543), 270-280.

Van Boven, L. and G. Loewenstein. 2003. "Social Projection of Transient Drive States." *Personality and Social Psychology Bulletin,* 29, 1159-1168.

Zettelmeyer, F., F. Scott Morton, J. Silva-Risso. 2007. "Scarcity Rents in Car Retailing: Evidence from Inventory Fluctuations at Dealerships." Mimeo, Northwestern University.

Figure 1 - Seasonal Trends in Vehicle Purchases. This figure illustrates the percentage of total vehicles that were sold in each month between 2001 and 2008 that were convertibles (Panel A) and 4-wheel drives (Panel B).

Panel A. Convertible Percentage of Vehicles Sold



Panel B. 4-Wheel Drive Percentage of Vehicles Sold

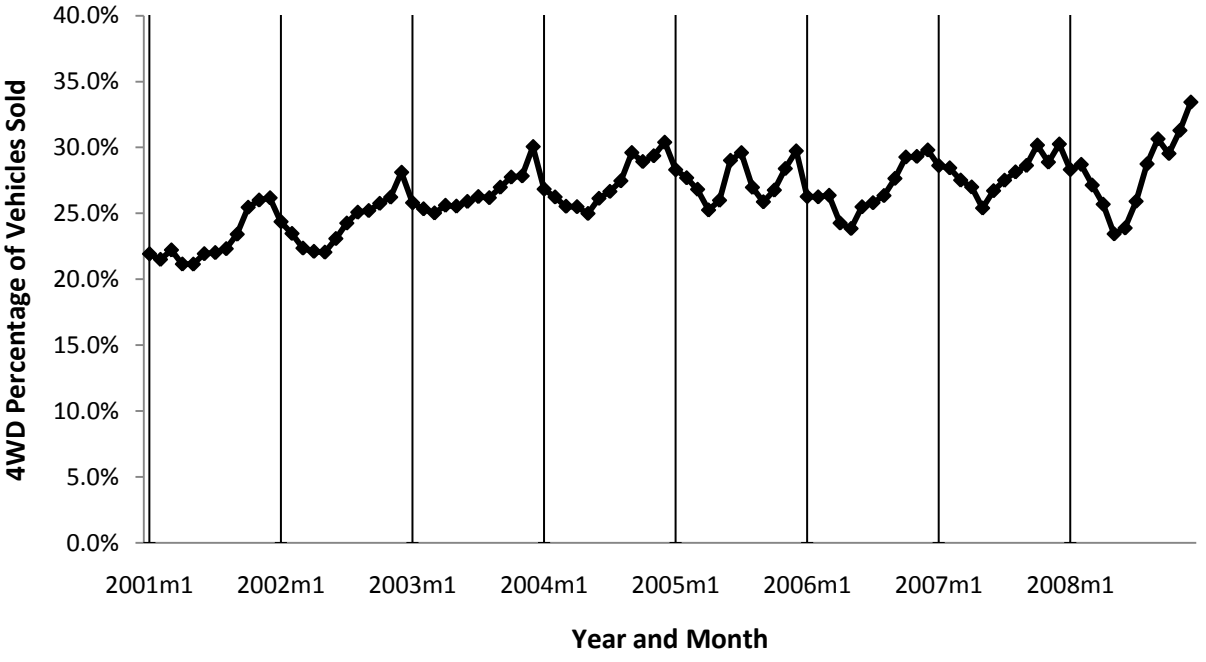
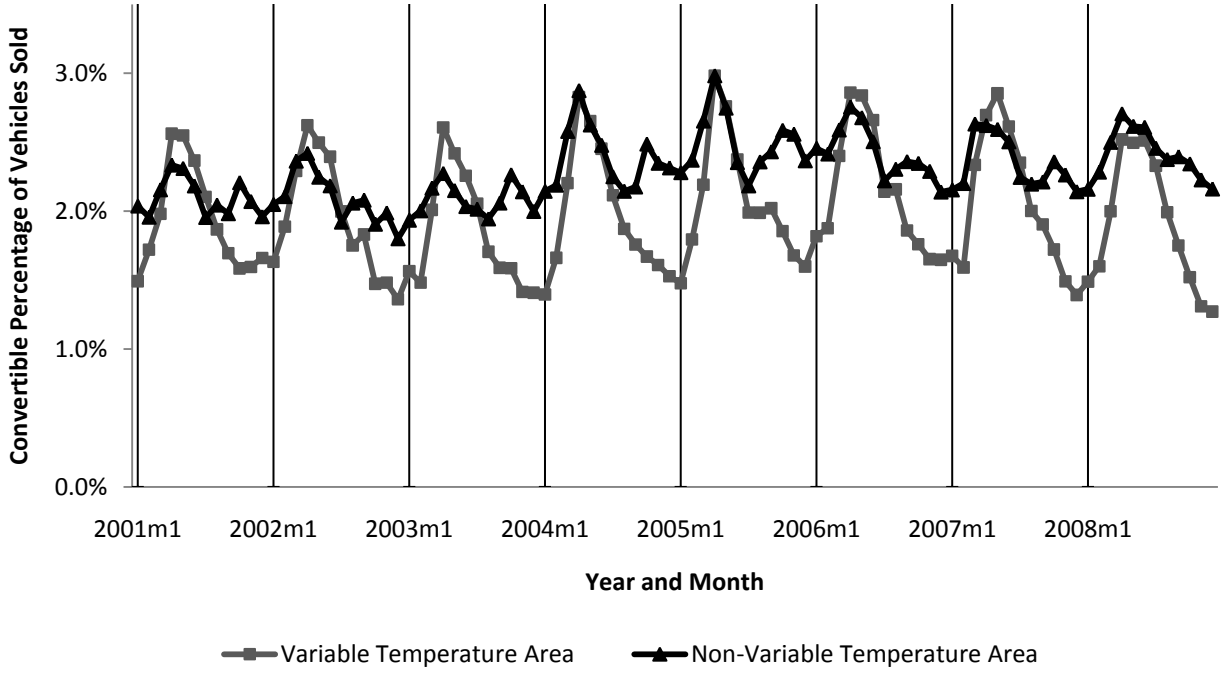


Figure 2 - Seasonal Trends in Vehicle Purchases by Temperature Variation. This figure illustrates the percentage of total vehicles sold between 2001 and 2008 that were convertibles (Panel A) and 4-wheel drives (Panel B) for DMAs with above- and below-median level of monthly DMA temeprature variation.

Panel A. Convertible Percentage of Vehicles Sold by Temperature Variation



Panel B. 4-Wheel Drive Percentage of Vehicles Sold by Temperature Variation

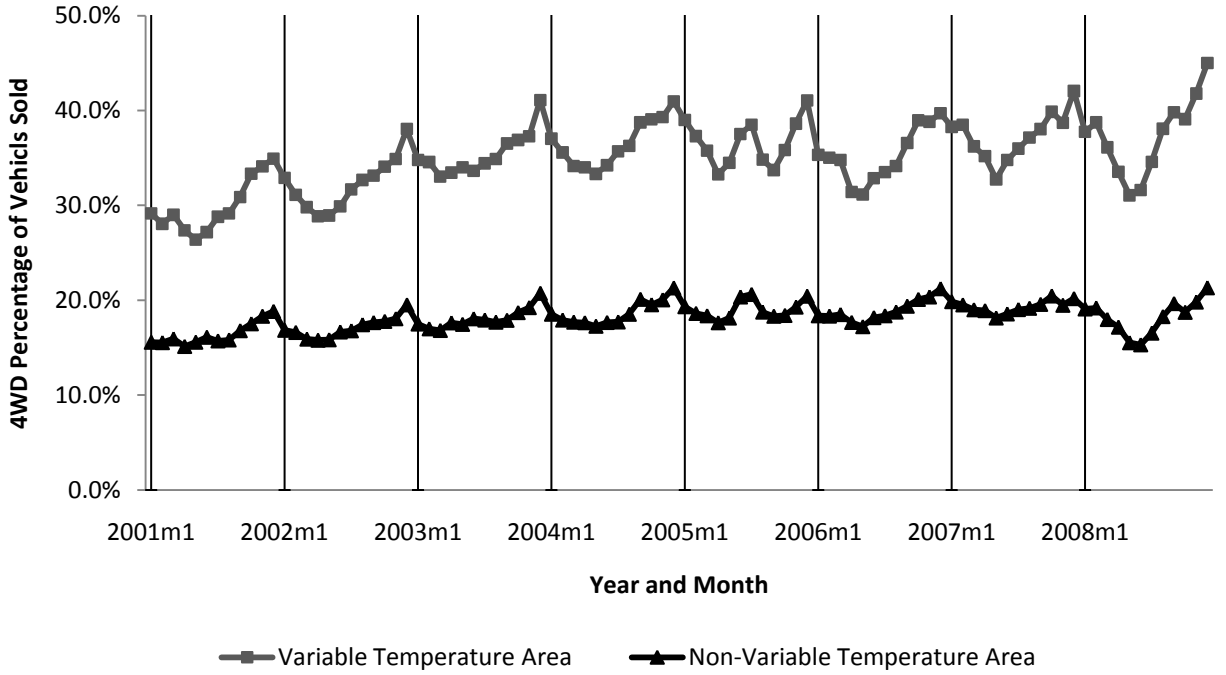
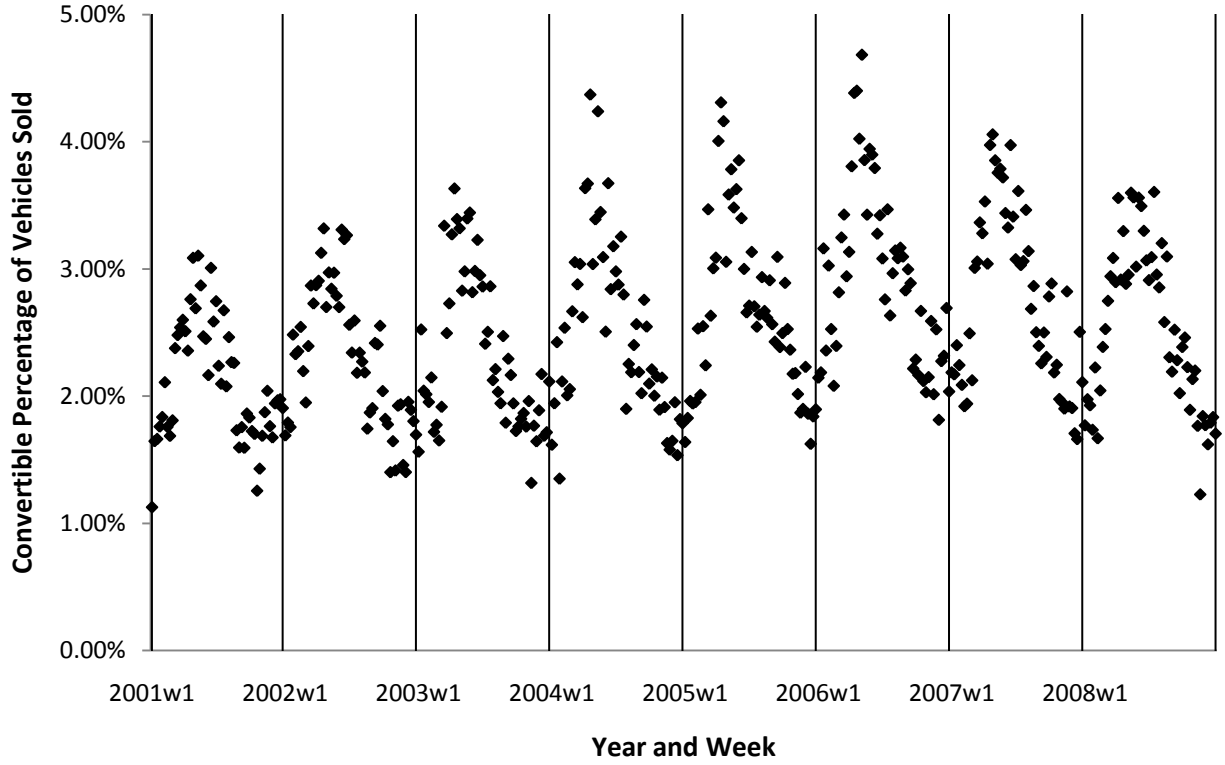


Figure 3. Convertible Sales - Chicago. Panel A illustrates the percentage of vehicles sold in Chicago for each of the 52 weeks in a year that were convertibles. Panel B plots the residual convertible percentage of vehicles sold in each week. (Residual is net of year and week-of-the-year fixed effects.)

Panel A. Convertible Percentage of Vehicles Sold - Chicago



Panel B. Residual Convertible Percentage of Vehicles Sold - Chicago

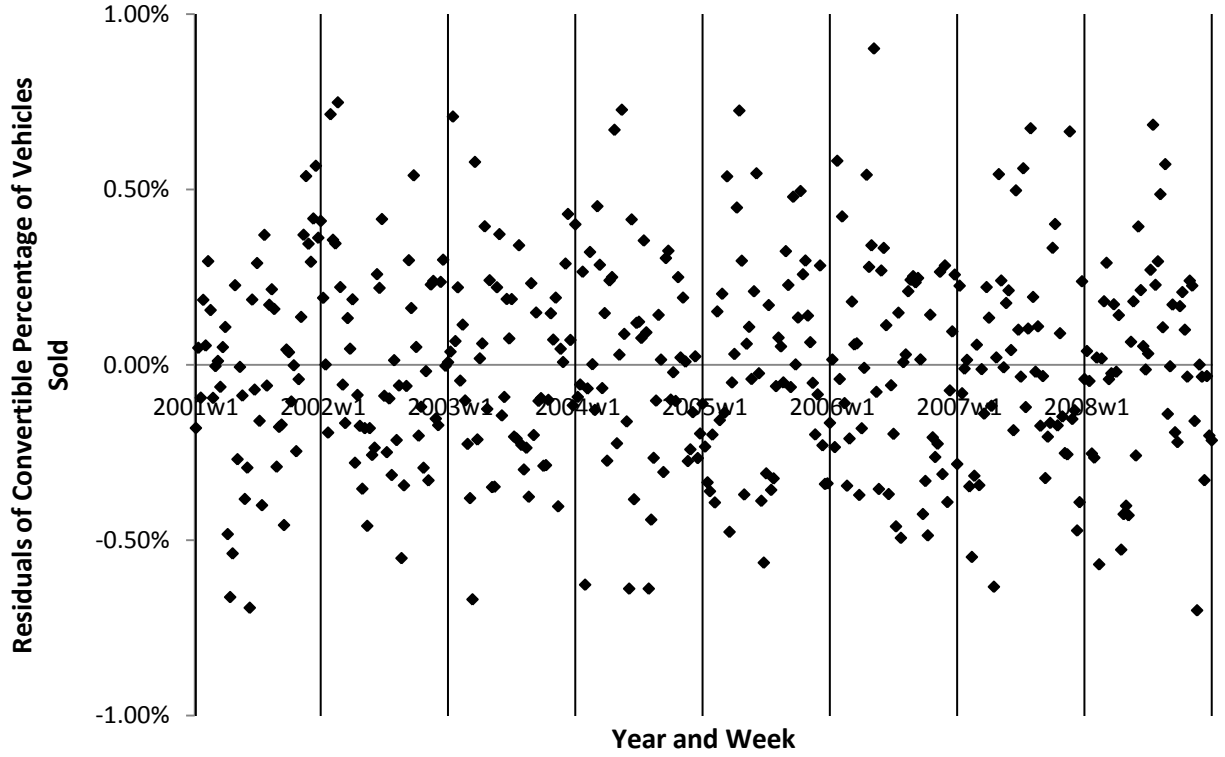
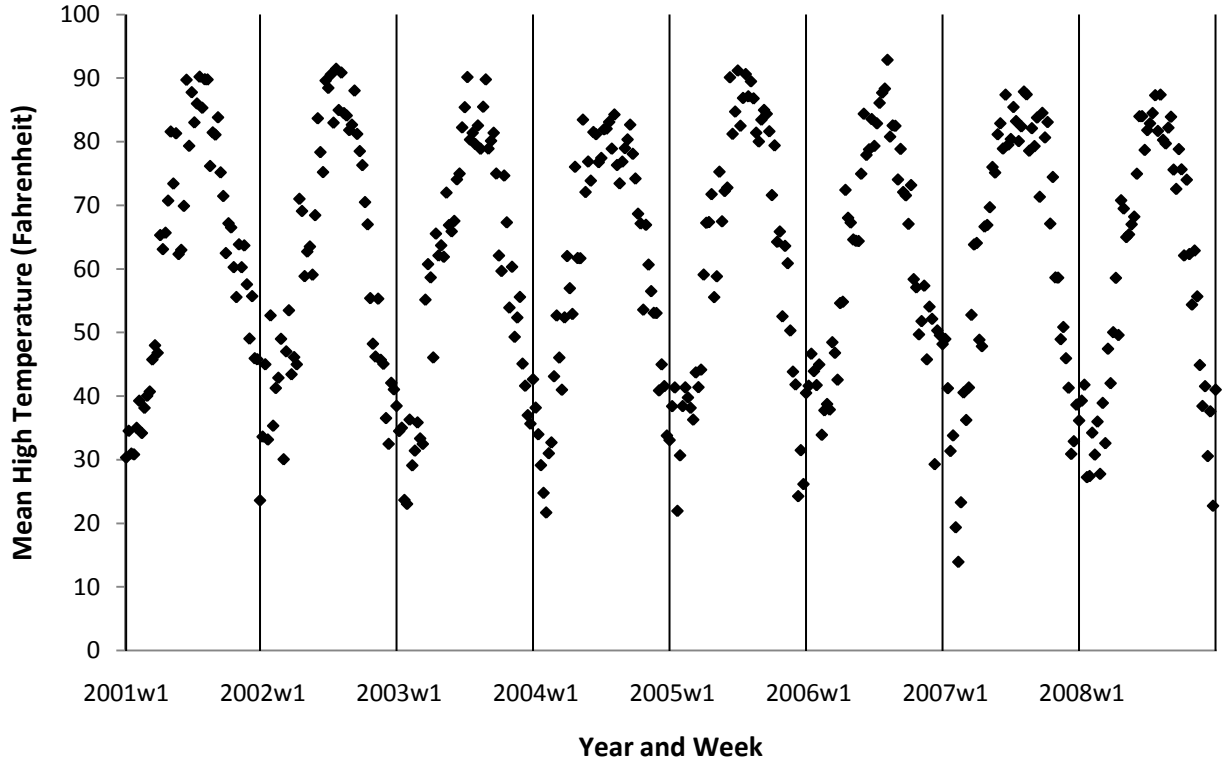


Figure 4. Temperature - Chicago. Panel A illustrates the average daily high temperature in Chicago for each of the 52 weeks in a year. Panel B plots the residual average daily high temperature in each week. (Residual is net of year and week-of-the-year fixed effects.)

Panel A. Mean High Temperature (Fahrenheit) - Chicago



Panel B. Residual Mean High Temperature (Fahrenheit) - Chicago

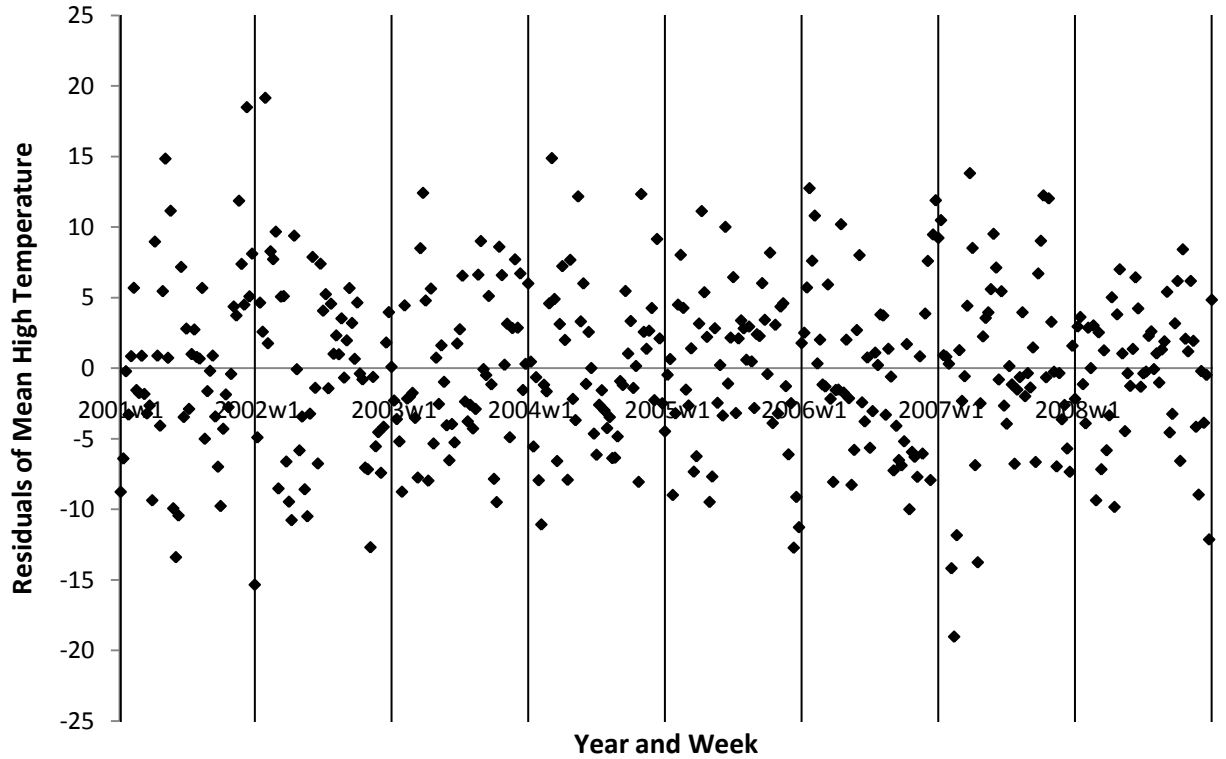
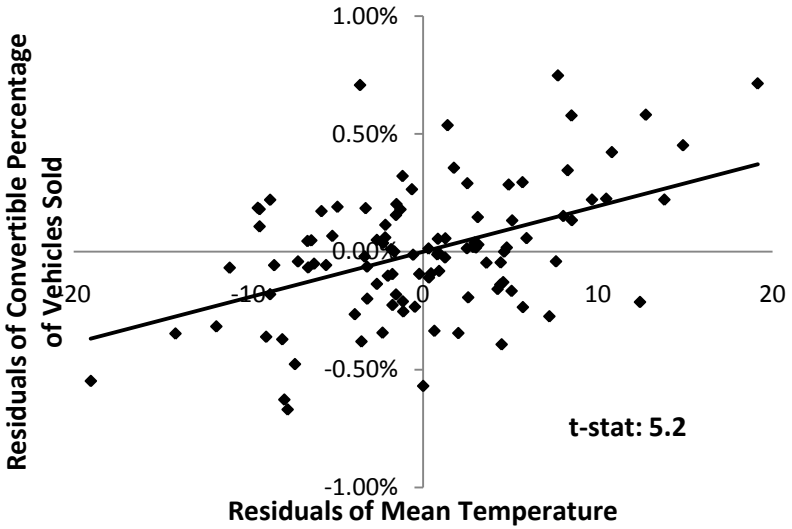
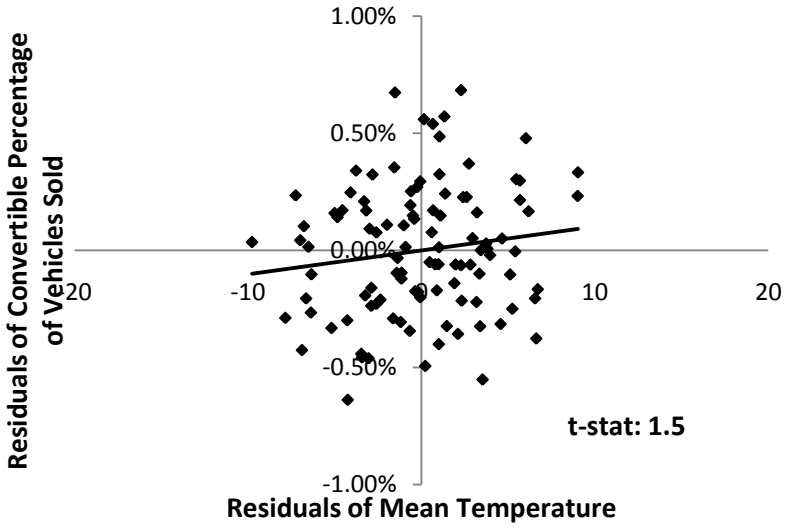


Figure 5. Temperature-Convertible Residuals - Chicago. This Figure provides scatter plots for the residuals of convertible percentage of vehicles sold (Panel B of Figure 3) and residuals of mean high temperature (Panel B of Figure 4) separately for each quarter of the year.

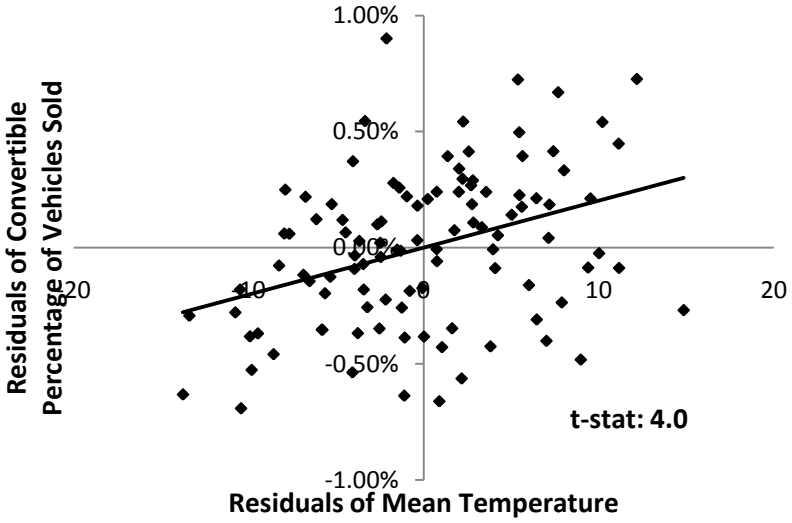
Panel A. Quarter 1



Panel C. Quarter 3



Panel B. Quarter 2



Panel D. Quarter 4

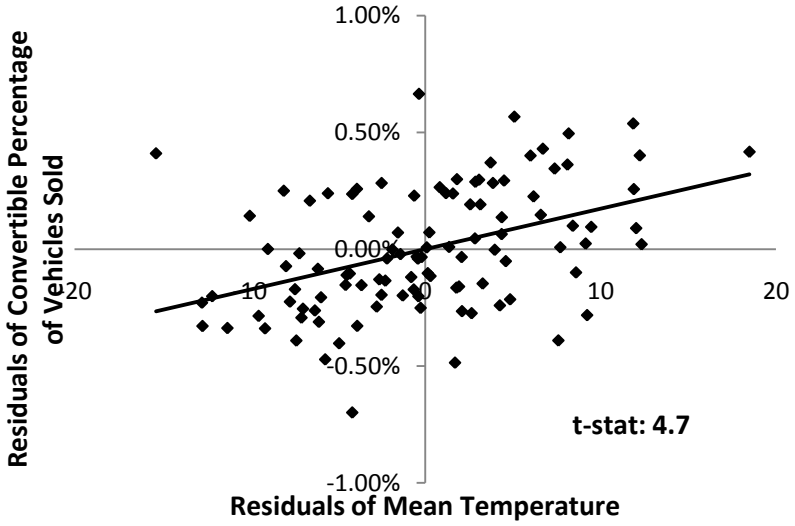
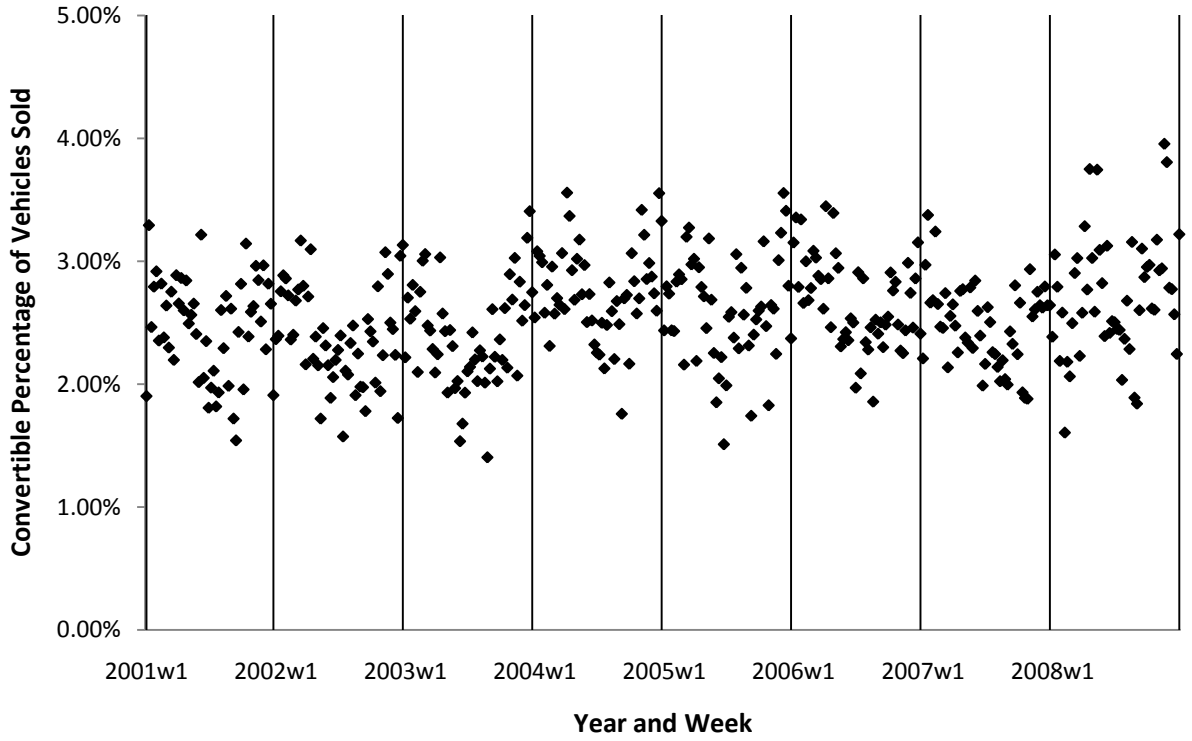


Figure 6. Convertible Sales - Miami. Panel A illustrates the percentage of vehicles sold in Miami-Ft. Lauderdale for each of the 52 weeks in a year that were convertibles. Panel B plots the residual convertible percentage of vehicles sold in each week. (Residual is net of year and week-of-the-year fixed effects.)

Panel A. Convertible Percentage of Vehicles Sold - Miami



Panel B. Residual Convertible Percentage of Vehicles Sold - Miami

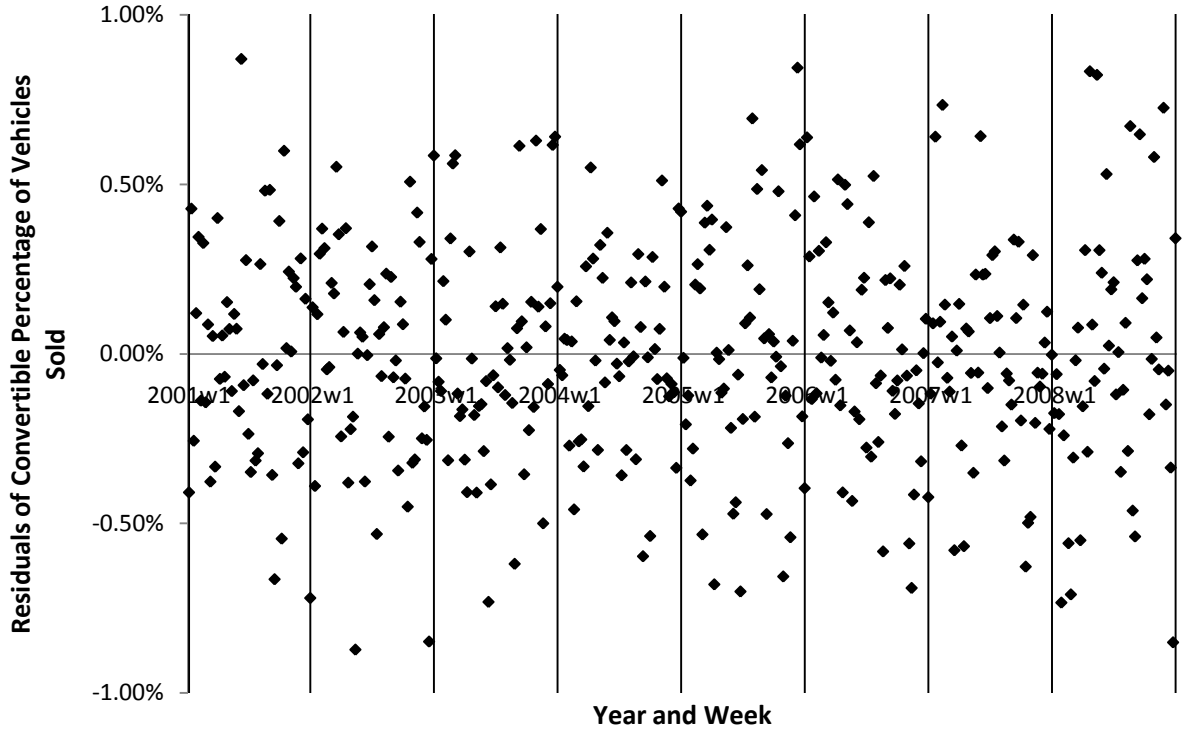
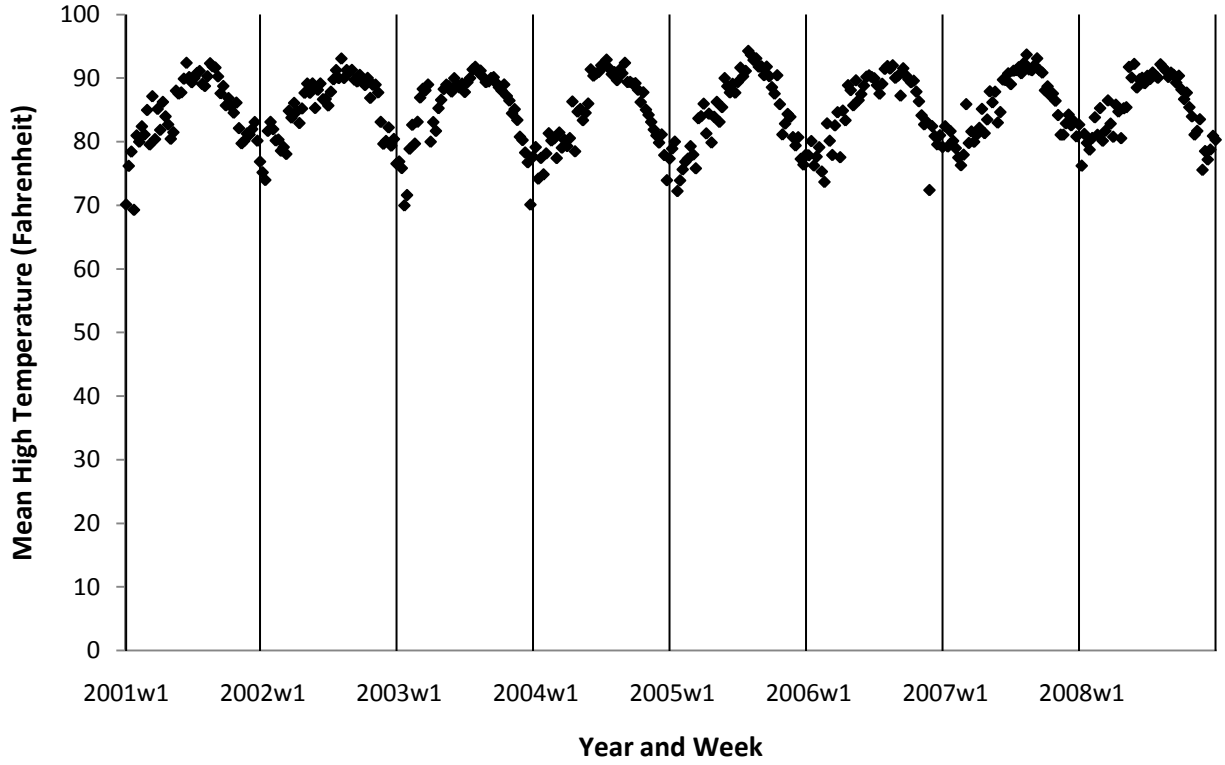


Figure 7. Temperature - Miami. Panel A illustrates the average daily high temperature in Miami-Ft. Lauderdale for each of the 52 weeks in a year. Panel B plots the residual average daily high temperature in each week. (Residual is net of year and week-of-the-year fixed effects.)

Panel A. Mean High Temperature (Fahrenheit) - Miami



Panel B. Residual Mean High Temperature (Fahrenheit) - Miami

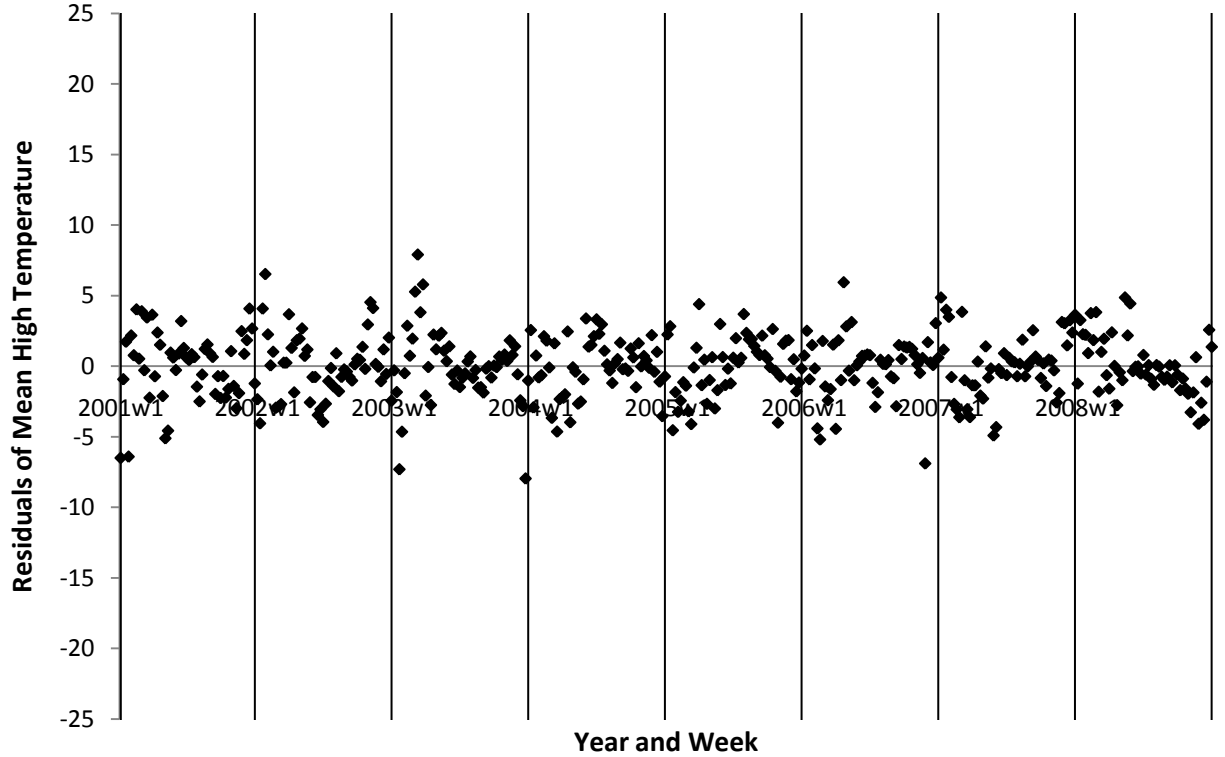
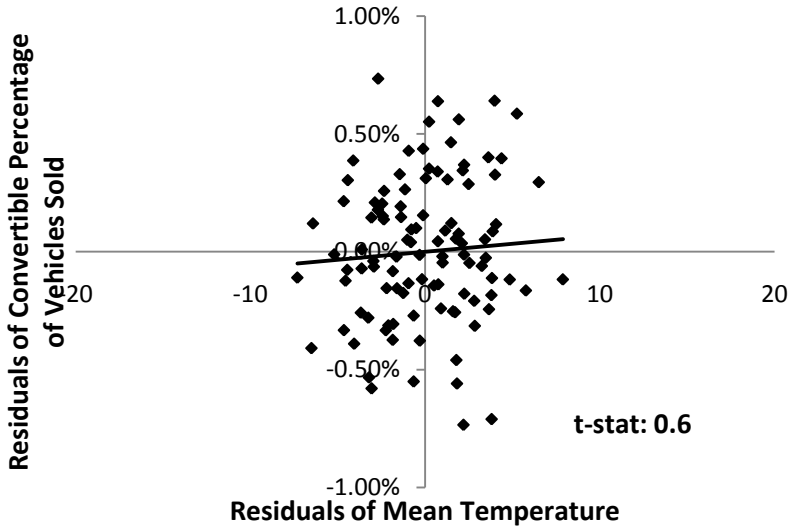
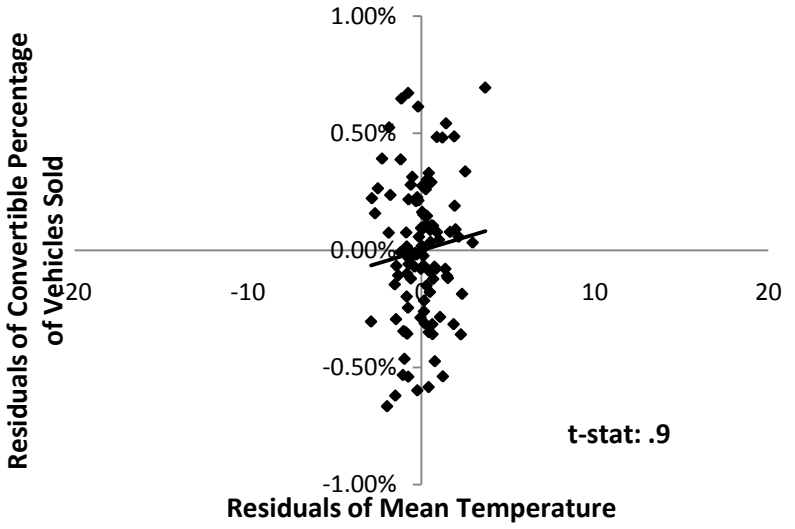


Figure 8. Temperature-Convertible Residuals - Miami. This Figure provides scatter plots for the residuals of convertible percentage of vehicles sold (Panel B of Figure 6) and residuals of mean temperature (Panel B of Figure 7) separately for each quarter of the year.

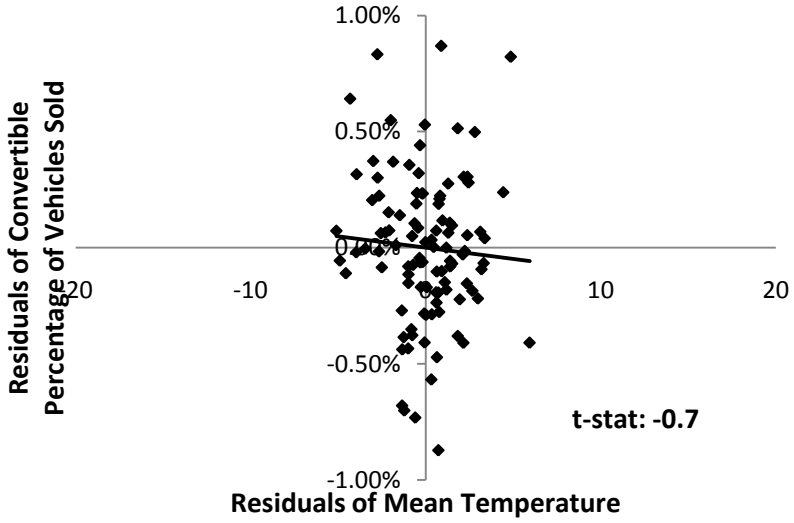
Panel A. Quarter 1



Panel C. Quarter 3



Panel B. Quarter 2



Panel D. Quarter 4

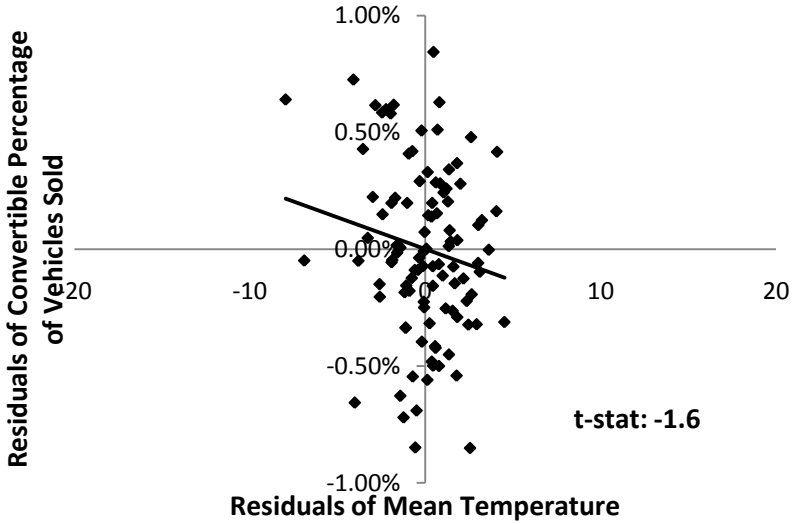


Figure 9. Temperature and Convertible Sales by Usual Temperature. This Figure provides the coefficient values and 95% confidence intervals for the impact of mean daily high temperature on convertible percentage of total vehicles sold (the estimate in Column 1 of Table 1) when the effect is estimated separately by the typical mean daily high temperature of the DMA-week-of-the-year. For example, the dot furthest to the left represents the estimated impact of temperature for DMA-weeks-of-the-year whose high temperature on average across the years in our sample was less than 35 degrees Fahrenheit.

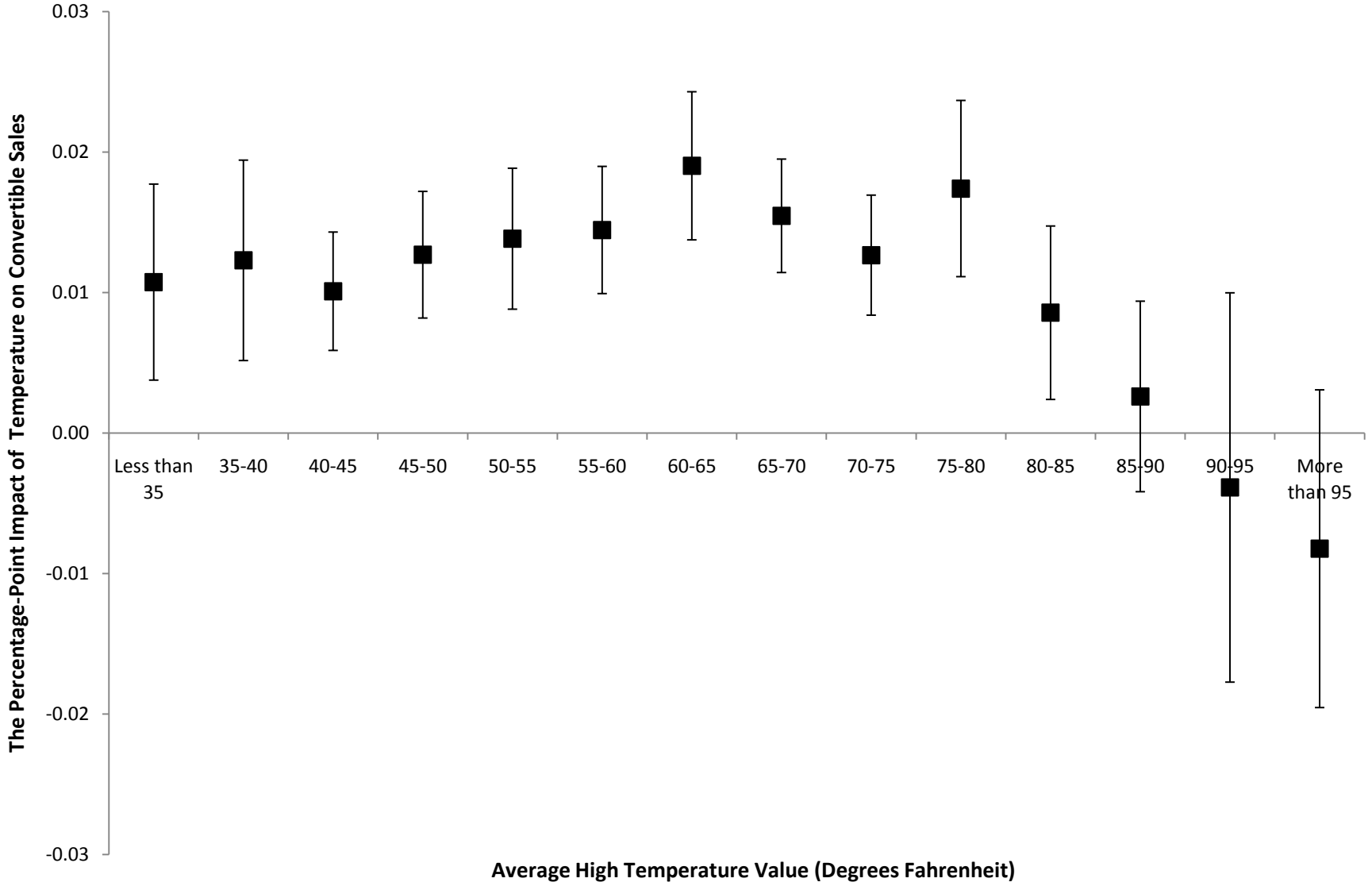


Figure 10. Snowfall and 4-Wheel Drive Sales - Event-Study Design. This Figure plots the weighted average and 95% confidence intervals for the residuals of the 4-wheel drive percentage of total vehicles sold for the twelve weeks leading up to and the twelve weeks after a snow storm event (week 0). The events were chosen to be the highest snow fall week of the year for DMAs that have above-median in weather variation.

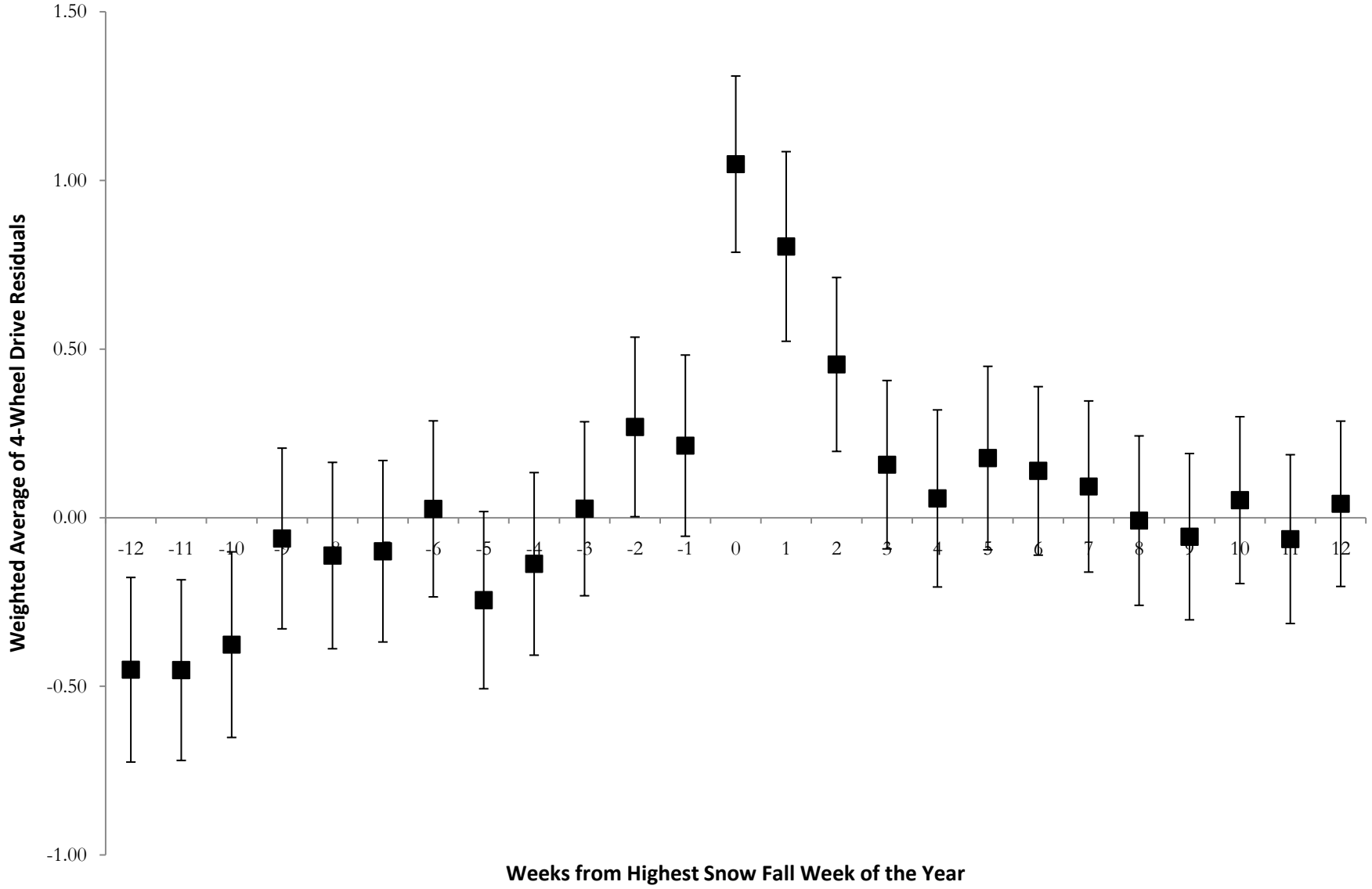
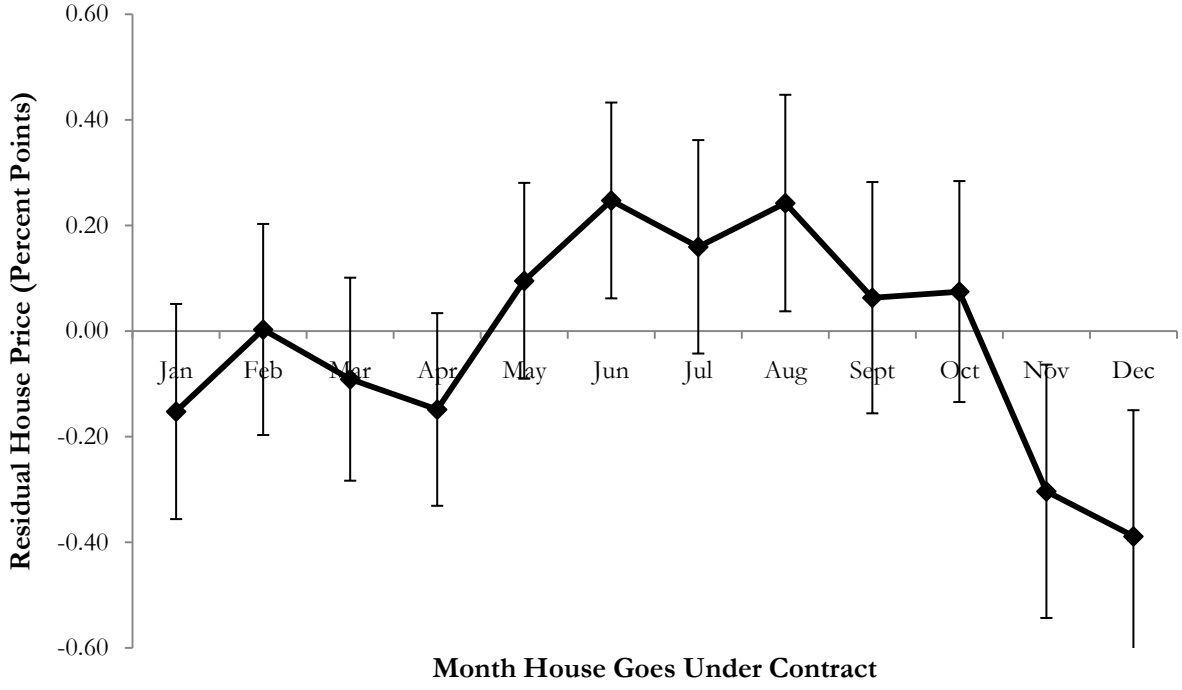


Figure 11 - Seasonal Value of a Swimming Pool. Panel A shows the average residual values for homes with swimming pools that go under contract during each month of the year. Panel B shows the estimated effect of a swimming pool on a house's residual sales price, conditional on other house characteristics, as estimated by Equation (7). 95% confidence intervals are also presented.

Panel A. Residuals by Month



Panel B. Conditional Effect of a Swimming Pool by Month

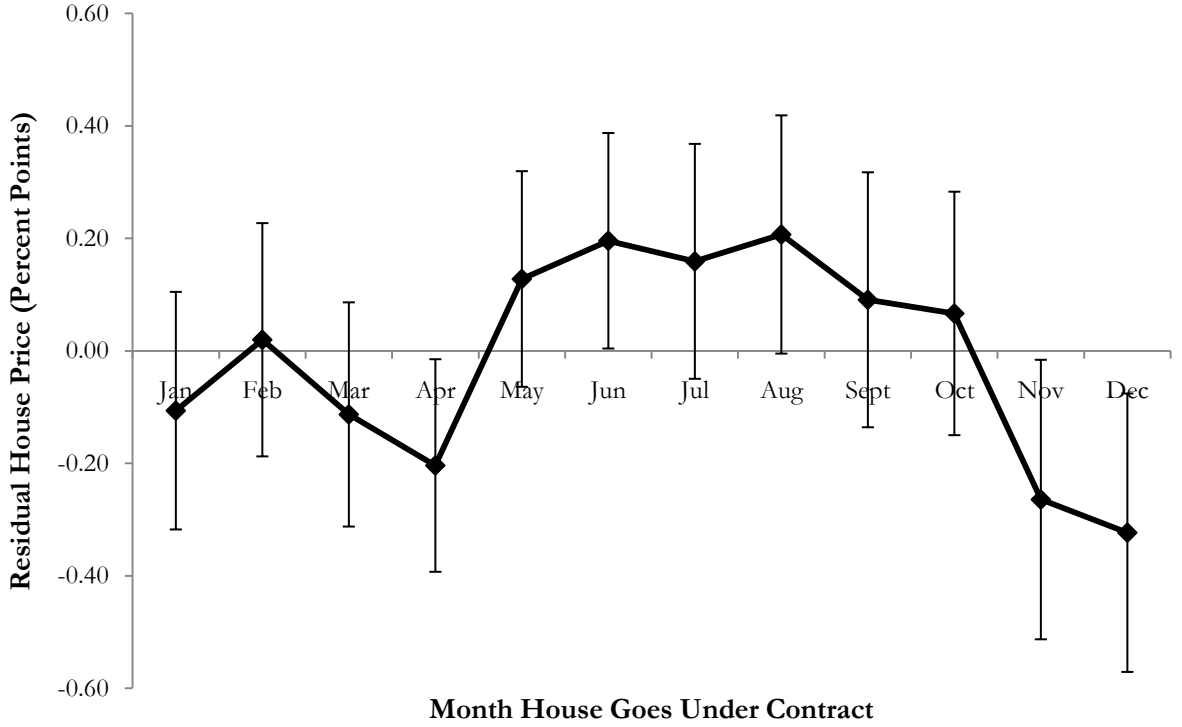
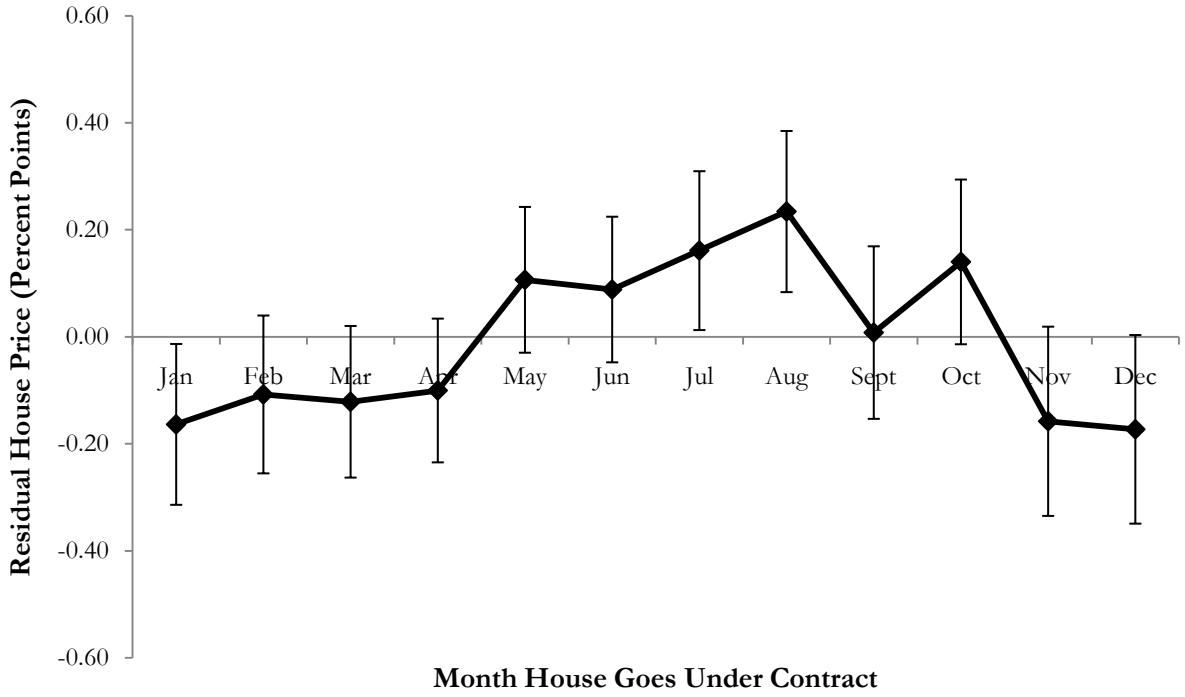
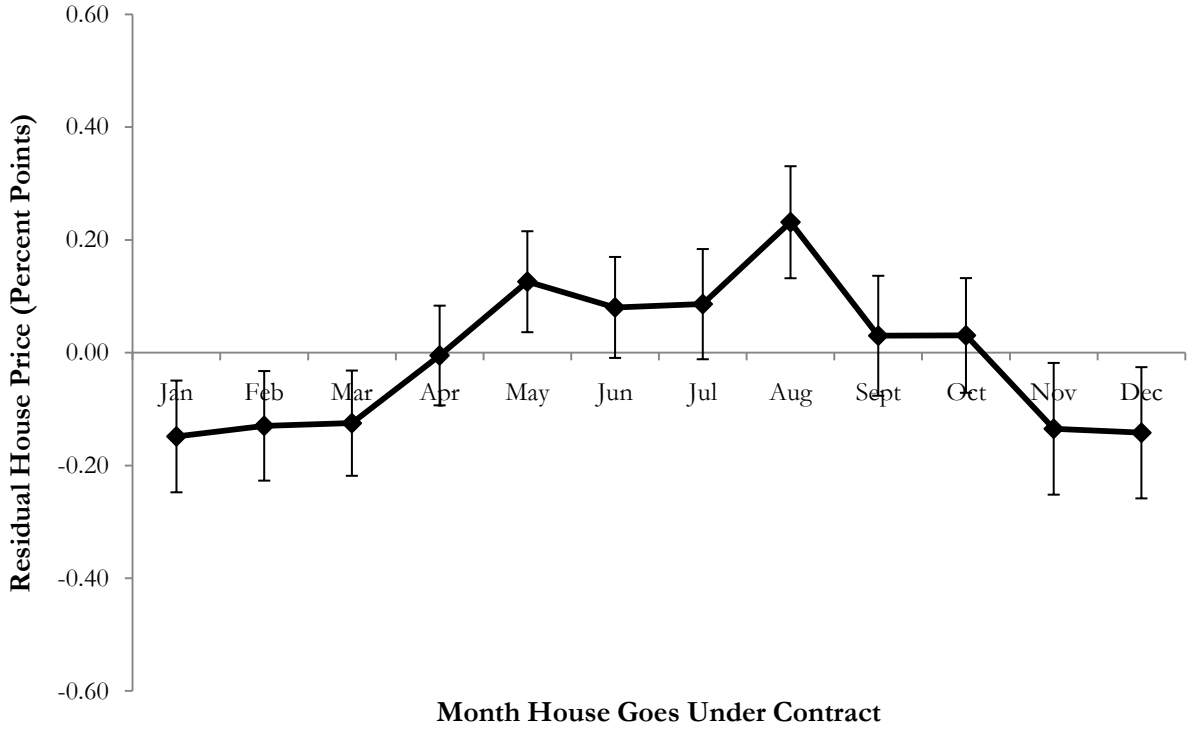


Figure 12 - Seasonal Value of a Swimming Pool - Trimming. Panel A provides the estimated effect of a swimming pool on a house's residual sales price, conditional on other house characteristics, as estimated by Equation (7), after eliminating residuals in the top and bottom 1%. Panel B shows the same estimated effects after eliminating residuals in the top and bottom 5%. 95% confidence intervals are also presented.

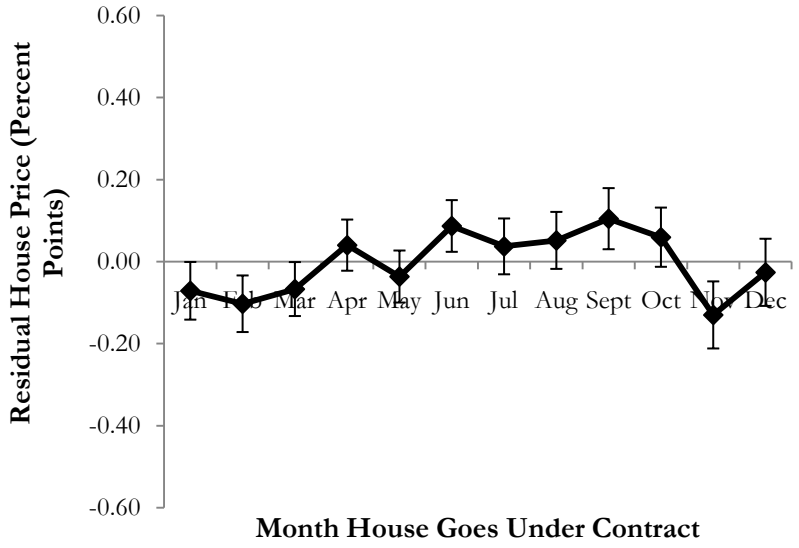
Panel A. Conditional Effect of a Swimming Pool by Month - 1% Trim



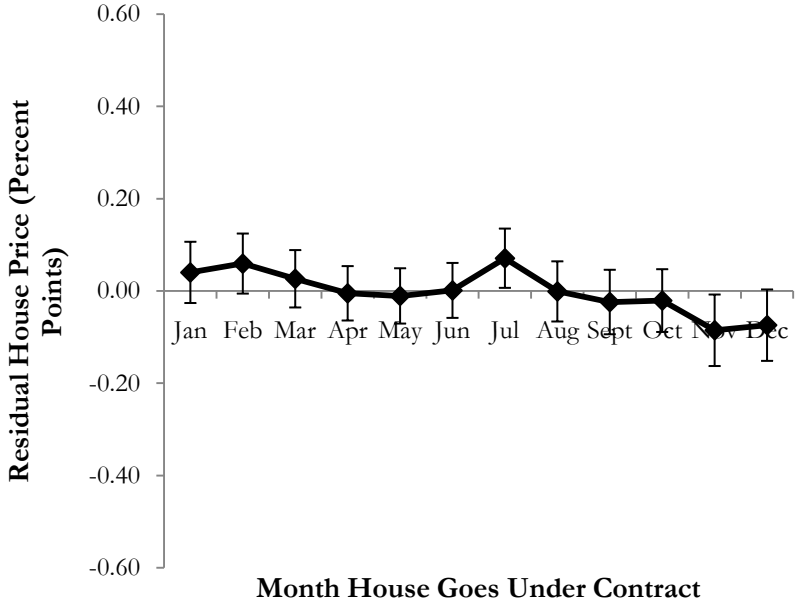
Panel B. Conditional Effect of a Swimming Pool by Month - 5% Trim



Panel A. Conditional Effect of Central Air by Month



Panel B. conditional Effect of a Fireplace by Month



Panel C. Conditional Effect of Lot Size by Month

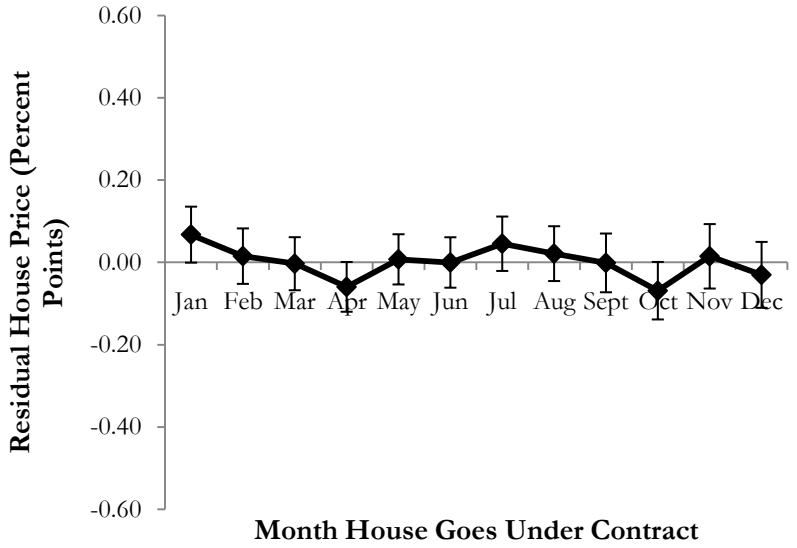
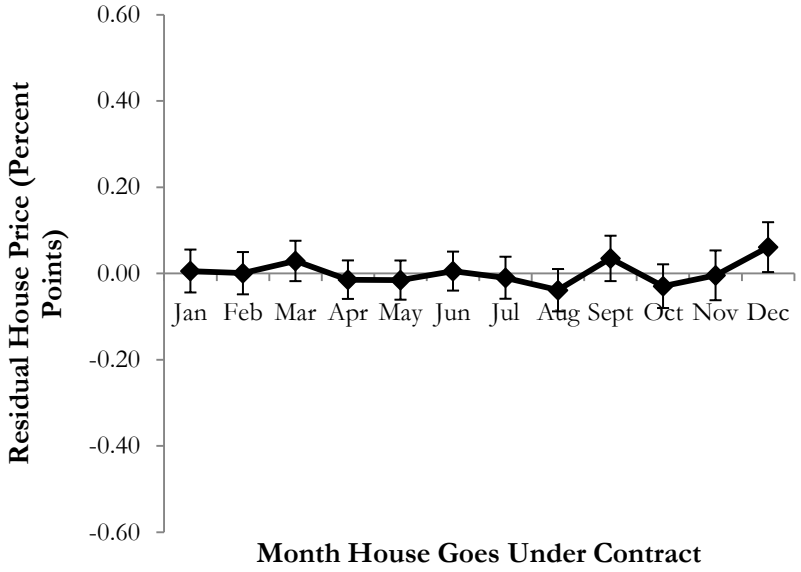
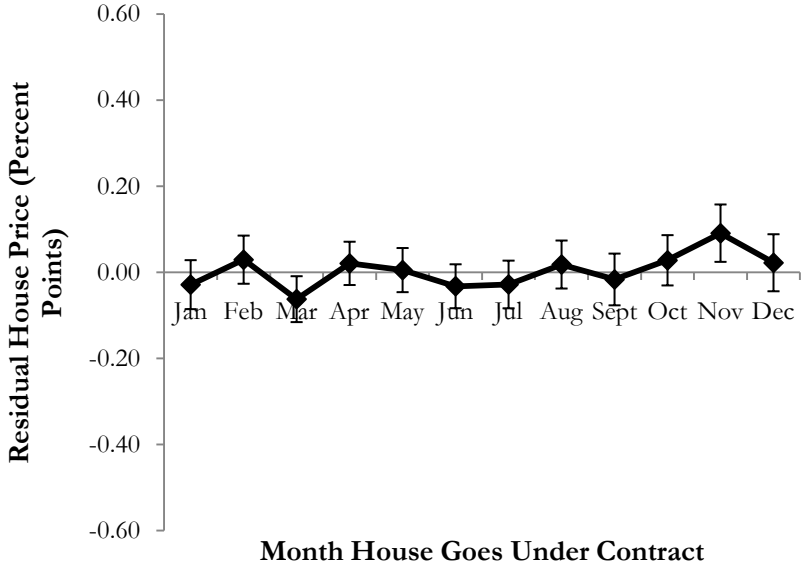


Figure 13 - Seasonal Value of Other Seasonal Housing Characteristics. This figure provides the estimated effect of a various characteristics on a house's residual sales price, conditional on other house characteristics, as estimated by Equation (7). Panel A shows the effect of central air, Panel B the effect of a fireplace, and Panel C the effect of lot size. The top and bottom 5% of residuals are removed. 95% confidence intervals are also presented.

Panel A. Conditional Effect of Number of Bedrooms by Month



Panel B. Conditional Effect of Number of Bathrooms by Month



Panel C. Conditional Effect of Square Footage by Month

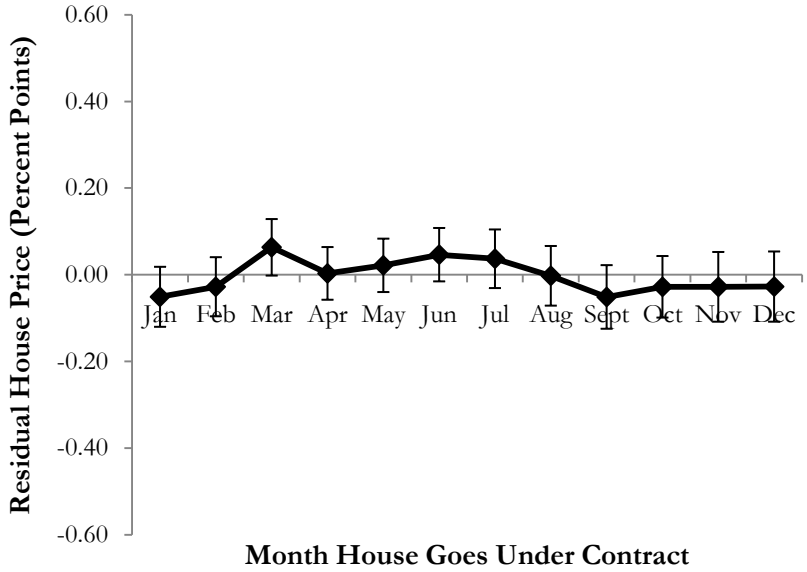


Figure 14 - Seasonal Value of Non-Seasonal Housing Characteristics. This figure provides the estimated effect of a various characteristics on a house's residual sales price, conditional on other house characteristics, as estimated by Equation (7). Panel A shows the effect of number of bedrooms, Panel B the effect of a number of bathrooms, and Panel C the effect of square footage. The top and bottom 5% of residuals are removed. 95% confidence intervals are also presented.

Table 1. Impact of Weather on Convertible Purchases

	Dep. Var.: Convertible Percentage of Total Vehicles Sold				
	Full Year	**Quarter 1**	**Quarter 2**	**Quarter 3**	**Quarter 4**
Temperature	.011**	.014**	.010**	.002	.011**
	(.000)	(.001)	(.002)	(.002)	(.001)
Rain Fall	-.005**	-.017**	-.006	-.000	-.003
	(.002)	(.004)	(.003)	(.003)	(.003)
Snow Fall	-.022	-.006	-.082	-	-.034
	(.024)	(.032)	(.106)	-	(.034)
Slush Fall	-.028**	-.020	-.028	-.026	-.033
	(.009)	(.014)	(.020)	(.026)	(.018)
Cloud Cover	-.172**	-.125*	-.342**	-.171**	-.108*
	(.027)	(.053)	(.057)	(.052)	(.044)
DMA*Year F.E.s	X	X	X	X	X
DMA*Week-of-the-Year F.E.s	X	X	X	X	X
R-Squared	0.778	0.837	0.780	0.813	0.860
Observations	49,499	11,637	13,123	12,798	11,941

Notes: Coefficient values and standard errors are presented from OLS regressions of the convertible percentage of total vehicles sold on weather variables - temperature (degrees Fahrenheit), rain (inches), snow (liquidized inches), slush (liquidized inches), and cloud cover (fraction of sky covered). Each observation is a DMA-Week and is weighted by the total number of vehicles sold. Fixed effects are included for DMA*Year and for DMA*Week-of-the-Year (Week 1 - Week 52). The first column uses all the data while the next four columns present results separately for the four quarters of the year.
* significant at 5%; ** significant at 1%

Table 2. Impact of Weather on 4-Wheel Drive Purchases

	Dep. Var.: 4-Wheel Drive Percentage of Total Vehicles Sold				
	Full Year	**Quarter 1**	**Quarter 2**	**Quarter 3**	**Quarter 4**
Temperature	-.050**	-.069**	-.024**	-.039**	-.063**
	(.002)	(.003)	(.004)	(.007)	(.004)
Rain Fall	.014*	.021	.029**	.031**	-.003
	(.006)	(.014)	(.010)	(.010)	(.012)
Snow Fall	1.02**	.73**	-.18	-8.11	1.18**
	(.05)	(.07)	(.20)	(25.3)	(.08)
Slush Fall	.24**	.24**	.12*	-.14	.45**
	(.02)	(.04)	(.04)	(.09)	(.05)
Cloud Cover	.378**	.351*	1.030**	.265	.405*
	(.082)	(.155)	(.158)	(.186)	(.150)
DMA*Year F.E.s	X	X	X	X	X
DMA*Week-of-the-Year F.E.s	X	X	X	X	X
R-Squared	0.964	0.972	0.971	0.970	0.972
Observations	68,431	16,517	17,101	17,320	17,493

Notes: Coefficient values and standard errors are presented from OLS regressions of the 4-wheel-drive percentage of total vehicles sold on weather variables - temperature (degrees Fahrenheit), rain (inches), snow (liquidized inches), slush (liquidized inches), and cloud cover (fraction of sky covered). Each observation is a DMA-Week and is weighted by the total number of vehicles sold. Fixed effects are included for DMA*Year and for DMA*Week-of-the-Year (Week 1 - Week 52). The first column uses all the data while the next four columns present results separately for the four quarters of the year.

* significant at 5%; ** significant at 1%

Table 3. Impact of Weather on Black Vehicle Purchases

	Dep. Var.: Black Percentage of Total Vehicles Sold				
	Full Year	**Quarter 1**	**Quarter 2**	**Quarter 3**	**Quarter 4**
Temperature	-.013**	-.012**	-.018**	-.006	-.010**
	(.001)	(.002)	(.003)	(.004)	(.002)
Rain Fall	.002	-.003	-.004	-.002	.011
	(.003)	(.007)	(.004)	(.005)	(.006)
Snow Fall	.097**	.102*	.273	-.145	.087
	(.032)	(.045)	(.151)	(.348)	(.050)
Slush Fall	.013	.021	.058*	.049	.013
	(.013)	(.020)	(.028)	(.044)	(.027)
Cloud Cover	.71**	.65**	.67**	.93**	.65**
	(.04)	(.09)	(.09)	(.094)	(.08)
DMA*Year F.E.s	X	X	X	X	X
DMA*Week-of-the-Year F.E.s	X	X	X	X	X
R-Squared	0.812	0.822	0.834	0.842	0.835
Observations	66,219	15,940	16,601	16,803	16,875

Notes: Coefficient values and standard errors are presented from OLS regressions of the percentage of total vehicles sold that are black in color on weather variables - temperature (degrees Fahrenheit), rain (inches), snow (liquidized inches), slush (liquidized inches), and cloud cover (fraction of sky covered). Each observation is a DMA*Week and is weighted by the total number of vehicles sold. Fixed effects are included for DMA*Year and for DMA*Week-of-the-Year (Week 1 - Week 52). The first column uses all the data while the next four columns present results separately for the four quarters of the year.
* significant at 5%; ** significant at 1%

Table 4. Impact of Weather on Convertible Purchases - Dynamic Analysis

	Dep. Var.: Convertible Percentage of Total Vehicles Sold				
	Full Year	**Quarter 1**	**Quarter 2**	**Quarter 3**	**Quarter 4**
Temperature Lead 1	.001	.003	.001	.005*	-.000
	(.001)	(.002)	(.002)	(.002)	(.002)
Temperature	.011**	.015**	.006**	.004	.010**
	(.001)	(.002)	(.002)	(.002)	(.002)
Temperature Lag 1	.001	.005**	-.005*	.001	.003
	(.001)	(.002)	(.002)	(.002)	(.002)
Temperature Lag 2	.003	.007**	-.001	.003	.000
	(.001)	(.002)	(.002)	(.002)	(.002)
Temperature Lag 3	.001	.002	-.001	.007**	-.001
	(.001)	(.002)	(.002)	(.002)	(.002)
Temperature Lag 4	.001	.002	-.003	.005	.001
	(.001)	(.002)	(.002)	(.002)	(.002)
Temperature Lag 5	-.001	.000	-.003	.003	-.001
	(.001)	(.002)	(.002)	(.002)	(.002)
Temperature Lag 6	.002*	-.001	.001	.010**	.004
	(.001)	(.002)	(.002)	(.002)	(.002)
Temperature Lag 7	.002*	.004**	.001	.002	-.003
	(.001)	(.002)	(.002)	(.002)	(.002)
Temperature Lag 8	.004**	.000	.006**	.008**	.003
	(.001)	(.002)	(.002)	(.002)	(.002)
Temperature Lag 9	.003**	.002	.000	.004	.004
	(.001)	(.002)	(.002)	(.002)	(.002)
Temperature Lag 10	-.000	.001	-.002	.004*	-.003
	(.001)	(.002)	(.001)	(.002)	(.003)
Temperature Lag 11	.000	-.002	-.000	.004	.007*
	(.001)	(.002)	(.001)	(.002)	(.003)
Temperature Lag 12	.000	.001	-.004**	.005*	.000
	(.001)	(.002)	(.001)	(.002)	(.003)
DMA*Year F.E.s	X	X	X	X	X
DMA*Week-of-the-Year F.E.s	X	X	X	X	X
Rain Fall (with Lead and Lags)	X	X	X	X	X
Snow Fall (with Lead and Lags)	X	X	X	X	X
Slush Fall (with Lead and Lags)	X	X	X	X	X
Cloud Cover (with Lead and Lags)	X	X	X	X	X
R-Squared	0.809	0.875	0.790	0.791	0.873
Observations	36,873	8,068	9,696	9,908	9,201

Notes: Coefficient values and standard errors are presented from OLS regressions of the convertible percentage of total vehicles sold on weather variables - temperature (degrees Fahrenheit), rain (inches), snow (liquidized inches), slush (liquidized inches), and cloud cover (fraction of sky covered). Both the current weather as well as the lead and 12 weekly lag weather variables are included in each regression. The coefficient values for rain, snow, slush, and cloud cover are omitted due to space constraints. Each observation is a DMA-Week and is weighted by the total number of vehicles sold. Fixed effects are included for DMA*Year and for DMA*Week-of-the-Year (Week 1 - Week 52). The first column uses all the data while the next four columns present results separately for the four quarters of the year.

* significant at 5%; ** significant at 1%

Table 5. Impact of Weather on 4-Wheel Drive Purchases - Dynamic Analysis

	Dep. Var.: 4-Wheel Drive Percentage of Total Vehicles Sold				
	Full Year	**Quarter 1**	**Quarter 2**	**Quarter 3**	**Quarter 4**
Snow Fall Lead 1	.18**	-.01	-.90	3.22**	.27**
	(.06)	(.06)	(.29)	(1.2)	(.09)
Snow Fall	1.01**	.71**	-.19	22.9	1.22**
	(.06)	(.09)	(.24)	(81.3)	(.10)
Snow Fall Lag 1	.85**	.60**	.04	21.6	1.29**
	(.06)	(.08)	(.23)	(121.5)	(.11)
Snow Fall Lag 2	.26**	.19*	.37	-247.8	.74**
	(.06)	(.07)	(.21)	(193.4)	(.13)
Snow Fall Lag 3	.12	.26**	-.04	147.4	.40**
	(.06)	(.08)	(.15)	(432.5)	(.14)
Snow Fall Lag 4	-.09	.01	-.39**	-250.9*	.26
	(.07)	(.08)	(.14)	(117.2)	(.17)
Snow Fall Lag 5	-.14*	-.07	-.35**	-156.3	.05
	(.07)	(.08)	(.12)	(116.2)	(.22)
Snow Fall Lag 6	-.26**	-.17	-.31**	-320.3**	-.07
	(.06)	(.08)	(.10)	(105.6)	(.39)
Snow Fall Lag 7	-.15*	-.15	-.05	-71.1*	-.59
	(.06)	(.08)	(.09)	(34.8)	(.52)
Snow Fall Lag 8	-.09	-.02	-.12	-39.4	1.58*
	(.06)	(.09)	(.09)	(32.6)	(.65)
Snow Fall Lag 9	-.09	-.05	-.11	-1.4	1.65*
	(.06)	(.09)	(.08)	(1.9)	(.74)
Snow Fall Lag 10	-.09	-.18	-.12	1.4	1.23
	(.06)	(.09)	(.08)	(.87)	(.77)
Snow Fall Lag 11	-.09	.05	-.25**	.01	-.33
	(.06)	(.10)	(.07)	(.75)	(.76)
Snow Fall Lag 12	-.13*	-.01	-.30**	.41	-1.68
	(.06)	(.09)	(.08)	(.29)	(1.2)
DMA*Year F.E.s	X	X	X	X	X
DMA*Week-of-the-Year F.E.s	X	X	X	X	X
Temperature (with Lead and Lags)	X	X	X	X	X
Rain Fall (with Lead and Lags)	X	X	X	X	X
Slush Fall (with Lead and Lags)	X	X	X	X	X
Cloud Cover (with Lead and Lags)	X	X	X	X	X
R-Squared	0.970	0.979	0.975	0.973	0.977
Observations	46,452	10,258	11,433	12,356	12,405

Notes: Coefficient values and standard errors are presented from OLS regressions of the 4-wheel drive percentage of total vehicles sold on weather variables - temperature (degrees Fahrenheit), rain (inches), snow (liquidized inches), slush (liquidized inches), and cloud cover (fraction of sky covered). Both the current weather as well as the lead and 12 weekly lag weather variables are included in each regression. The coefficient values for temperature, rain, slush, and cloud cover are omitted due to space constraints. Each observation is a DMA-Week and is weighted by the total number of vehicles sold. Fixed effects are included for DMA*Year and for DMA*Week-of-the-Year (Week 1 - Week 52). The first column uses all the data while the next four columns present results separately for the four quarters of the year.

* significant at 5%; ** significant at 1%

Table 6. Impact of Weather on Convertible and 4-Wheel Drive Purchases for Consumers Trading in a Convertible or 4-Wheel Drive Vehicle, Respectively

	Dep. Var.: Convertible or 4-Wheel Drive Percentage of Total Vehicles Sold	
	Convertibles	**4-Wheel Drives**
Temperature	.060**	-.044**
	(.020)	(.004)
Rain Fall	.007	.003
	(.042)	(.011)
Snow Fall	-.23	.61**
	(.60)	(.09)
Slush Fall	-.45	.19**
	(.24)	(.04)
Cloud Cover	-1.26	.89**
	(.679)	(.15)
DMA*Year F.E.s	X	X
DMA*Week-of-the-Year F.E.s	X	X
R-Squared	0.675	0.815
Observations	23,529	65,356

Notes: Coefficient values and standard errors are presented from OLS regressions of the convertible percentage of total cars sold (Column 1) and the 4-wheel-drive percentage of total vehicles sold (Column 2) on weather variables - temperature (degrees Fahrenheit), rain (inches), snow (liquidized inches), slush (liquidized inches), and cloud cover (fraction of sky covered). Each observation is a DMA-Week and is weighted by the total number of vehicles sold. Fixed effects are included for DMA*Year and for DMA*Week-of-the-Year (Week 1 - Week 52). The sample is restricted to people who were purchasing a vehicle while trading in a convertible (Column 1) or a 4-wheel drive (Column 2).
* significant at 5%; ** significant at 1%

Table 7. Impact of Weather on Quickly Trading In a Vehicle

	Dep. Var.: Dummy Variable if Returned Within 1-3 Years		
	1 Year	**2 Years**	**3 Years**
Mean of Dependent Variable	2.37%	5.03%	7.16%
Convertible	1.272%**	2.302%**	2.905%**
	(.019%)	(.030%)	(.042%)
Convertible Interacted with:			
Temperature	.006%	.017%**	.006%
	(.004%)	(.007%)	(.009%)
Rain Fall	.008%	.002%	-.018%
	(.009%)	(.015%)	(.021%)
Snow Fall	.181%	-.041%	-.142%
	(.131%)	(.222%)	(.289%)
Slush Fall	.063%	.028%	-.116%
	(.053%)	(.094%)	(.131%)
Cloud Cover	-.197%	-.036%	.332%
	(.138%)	(.228%)	(.312%)
4-Wheel Drive	.285%**	.929%**	1.634%**
	(.006%)	(.006%)	(.014%)
4-Wheel Drive Interacted With:			
Temperature	-.003%*	-.005%*	-.013%**
	(.001%)	(.002%)	(.003%)
Rain Fall	-.005%	-.005%	.001%
	(.003%)	(.006%)	(.008%)
Snow Fall	.000%	.063%	.004%
	(.035%)	(.058%)	(.076%)
Slush Fall	.002%	-.019%	-.048%
	(.016%)	(.028%)	(.038%)
Cloud Cover	.006%	-.109%	-.124%
	(.047%)	(.078%)	(.106%)
DMA*Week Fixed Effects	X	X	X
R-Squared	0.004	0.006	0.007
Observations	35,102,062	29,665,047	23,827,418

Notes: Coefficient values and standard errors are presented from OLS regressions of a dummy variable for whether the vehicle shows up in our dataset (as a trade-in car or another car sale) within 1, 2, or 3 years from the date of purchase on a convertible and 4-wheel drive dummy variable and an interaction between these vehicle types and weather variables at the time of purchase - temperature (degrees Fahrenheit), rain (inches), snow (liquidized inches), slush (liquidized inches), and cloud cover (fraction of sky covered). Each observation is at the individual vehicle level and DMA*Week fixed effects are included. The dataset is also restricted so as to eliminate all truncation (Columns 1-3 eliminate the last 1-3 years of car sales in the sample, respectively).

* significant at 5%; ** significant at 1%

Table 8. Impact of Weather on Convertible and 4-Wheel Drive Purchase Price

	Dependent Variable: Vehicle Sales Price (Less Rebate)			
	Convertibles		4-Wheel Drives	
	New	Used	New	Used
Mean of Dependent Variable	$40,001	$22,222	$31,845	$19,132
Temperature	1.22	3.98**	.83*	2.03**
	(1.46)	(1.50)	(.38)	(.29)
Rain Fall	1.56	2.67	-.05	.74
	(3.16)	(3.40)	(.98)	(.80)
Snow Fall	50.68	-114.48*	3.33	-23.80**
	(46.02)	(46.45)	(8.77)	(6.91)
Slush Fall	-6.92	-30.39	-5.33	5.03
	(17.83)	(17.90)	(4.16)	(3.11)
Cloud Cover	25.24	-64.52	36.44**	-54.06**
	(49.69)	(51.68)	(13.96)	(11.04)
DMA*Year F.E.s	X	X	X	X
DMA*Week-of-the-Year F.E.s	X	X	X	X
Purchase Timing F.E.s	X	X	X	X
Vehicle-type F.E.s	X	X	X	X
Odometer Value Spline		X		X
Observations	391,438	377,321	5,495,657	4,152,489

Notes: Coefficient values and standard errors are presented from OLS regressions of vehicle transaction prices on weather variables - temperature (degrees Fahrenheit), rain (inches), snow (liquidized inches), slush (liquidized inches), and cloud cover (fraction of sky covered). Each observation is an individual transaction. Fixed effects are included for DMA*Year and for DMA*Week-of-the-Year (Week 1 - Week 52), and for detailed vehicle types. Purchase Timing indicates whether a vehicle was purchased on a weekend or at the end of the month. The first two columns present results for new and used convertibles, respectively, while the second two columns present results for new and used 4-wheel drives. The used vehicle specifications (columns 2 and 4) include an linear spline in odometer values with knots at 10,000 mile increments.

* significant at 5%; ** significant at 1%

Table 9. Summary Statistics for Retail Vehicle Sales, by DMA*Week

	Mean	St. Dev.	Min	Max
Vehicle Characteristics				
Number of Convertibles Sold	12.4	25.4	0	287
4-Wheel Drives Sold	153.0	292.3	0	6220
Total Vehicles Sold	574.9	1029.9	1	11633
Percentage Convertibles	1.8%	2.6%	0%	100%
Percentage 4-Wheel Drives	30.2%	18.5%	0%	100%
Percentage Black Vehicles	11.2%	6.2%	0%	100%
Weather Variables				
Temperature	70.1	18.3	-26.1	115.8
Rain Fall	1.4	2.4	0	52.9
Snow Fall	.04	.23	0	10.5
Slush Fall	.10	.48	0	22.0
Cloud Cover	.47	.23	0	1
Observations	70,790	70,790	70,790	70,790

Notes: Summary statistics reported for DMA*Week observations.

Table 10. Housing Summary Statistics

	Mean	Standard Deviation	Minimum	Maximum
Sales Price	273,925	263,681	5,001	5,000,000
Swimming Pool	0.119	0.321	0	1
Central Air	0.304	0.460	0	1
Fireplace	0.455	0.337	0	1
Lot Size (Acres)	0.320	0.487	0	5
Year Built	1968	24	1900	2006
Square Footage	1679	734	250	10000
Bathrooms	2.06	0.84	0.5	10
Bedrooms	3.12	0.81	1	10
Observations	4,206,314	4,206,314	4,206,314	4,206,314

Table 11. The Impact of Temperature and Housing Characteristics on Residual Sales Prices

	Dependent Variable: Residual Housing Prices							
	Linear Temperature		**Temperatre > 70° F**		**Temperatre > 80° F**		**Temperatre > 90° F**	
Interaction of Temperature and:								
Swimming Pool	.013**	.010**	.23**	.18**	.27**	.13**	.41	.37**
	(.004)	(.002)	(.06)	(.03)	(.09)	(.04)	(.30)	(.14)
Fire Place	.0006	.0013	-.01	.02	.06	.02	.18	-.12
	(.0024)	(.0011)	(.05)	(.02)	(.07)	(.04)	(.24)	(.11)
Lot Acre	.0006	.0006	-.06	-.05*	.00	-.02	-.68*	-.28
	(.0024)	(.0012)	(.05)	(.02)	(.09)	(.04)	(.29)	(.15)
Central Air	.0002	.0060**	.02	.10**	.02	.03	-.13	-.23
	(.0025)	(.0012)	(.05)	(.02)	(.07)	(.04)	(.25)	(.12)
Square Footage (1,000s)	.0043	.0004	.10*	.04	.09	-.03	.42	.22
	(.0025)	(.0012)	(.05)	(.02)	(.08)	(.04)	(.27)	(.13)
Number of Baths	-.0034	-.0008	-.03	-.03	-.05	.03	.07	.09
	(.0020)	(.0010)	(.04)	(.02)	(.07)	(.03)	(.25)	(.12)
Number of Bedrooms	.0004	-.0009	.01	-.02	.02	.00	-.34	-.09
	(.0018)	(.0009)	(.03)	(.02)	(.06)	(.03)	(.17)	(.08)
Levels of All Variables	X	X	X	X	X	X	X	X
Trim 5%		X		X		X		X
Observations	4,145,410	3,731,014	4,145,410	3,731,014	4,145,410	3,731,014	4,145,410	3,731,014

Notes: The first two columns of this table present coefficients and standard errors from the regression of residual housing prices (from Equation (6) in the text) on the interaction between housing characteristics and linear temperature (average high daily temperature during the month the house goes under contract). The next three sets of columns report the interaction between housing characteristics and dummy variables for the average daily high temperature in the month of housing contract being above 70, 80, or 90 degrees Fahrenheit. The second column in each set restricts the sample to house sales whose residuals were not in the top or bottom 5%. The level effects of all variables (not just the interactions) are also included in all of the regressions. All coefficients are multiplied by 100 to make them easier to read (see text). Thus, the coefficients can be interpretted as approximate percentage-point changes.

* significant at 5%; ** significant at 1%

Buy-it-now or Take-a-chance: Price Discrimination through Randomized Auctions *

L. Elisa Celis Gregory Lewis Markus M. Mobius
Hamid Nazerzadeh

July 3, 2012

Abstract

Online tracking technology allows platforms to offer advertisers targeted consumer demographics, improving match quality, but thinning markets. Bidding data from Microsoft exhibits a large gap in the top two bids, consistent with this intuition. This motivates our new mechanism. Bidders can "buy-it-now", or "take-a-chance" in an auction where the top $d > 1$ bidders are equally likely to win. The randomized allocation incentivizes high valuation bidders to buy-it-now. This mechanism dominates the second-price auction, and approximates Myerson's optimal mechanism. Running counterfactual simulations, we find it improves revenue by 4.5% and consumer surplus by 11% compared to an optimal second-price auction.

*Department of Computer Science, University of Washington (ecelis@cs.washington.edu), Department of Economics, Harvard University (glewis@fas.harvard.edu), Microsoft Research New England (mobius@microsoft.com), and USC Marshall School of Business (nazerzad@marshall.usc.edu) respectively. All authors are grateful to Microsoft Research New England for their hospitality, and Greg would also like to thank NET Institute (www.NETinst.org) for financial support. Josh Feng and Danyang Su provided excellent research assistance. We thank Susan Athey, Alessandro Bonatti, Michael Grubb, Mallesh Pai and Lee Zen for fruitful discussions that improved this work. Finally we thank Lee Zen and members of Microsoft's display advertisting team— in particular Ian Ferreira, Keith Hawley, and Brandon Zirkle — for their help.

1 Introduction

Advertising technology is changing fast. Consumers can now be reached while browsing the internet, playing games on their phone or watching videos on YouTube. The large companies that control these new media — household names like Google, Facebook and Yahoo! — generate a substantial part of their revenue by selling advertisements. They also know increasing amounts of information about their users. This allows them to match advertisers to potential buyers with ever greater efficiency. While this matching technology generates surplus for advertisers, it also tends to create thin markets where perhaps only a single advertiser has a high willingness to pay. These environments pose special challenges for the predominant auction mechanisms that are used to sell online ads because they reduce competition among bidders, making it difficult for the platform to extract the surplus generated by targeting.

For example, a sportswear firm advertising on the New York Times website may be willing to pay much more for an advertisement placed next to a sports article than one next to a movie review. It might pay an additional premium for a local consumer who lives in New York City and an even higher premium if the consumer is known to browse websites selling sportswear. Each layer of targeting increases the sportswear firm's valuation for the consumer but also dramatically narrows the set of participating bidders to fellow sportswear firms in New York City. Without competition, revenue performance may be poor (Bergemann and Bonatti 2010, Levin and Milgrom 2010).

Consider a simple model: When advertisers "match" with users, they have high valuation; otherwise they have low valuation. Assume that match probabilities are independent across bidders, and sufficiently low that the probability that *any* bidder matches is relatively small. Then a second-price auction will typically get low revenue, since the probability of two "matches" occurring in the same auction is small. On the other hand, setting a high fixed price is not effective since the probability of zero "matches" occurring is relatively large and many impressions would go unallocated. Hence, allowing targeting creates asymmetries in valuations that can increase efficiency, but decrease revenue. In fact, because of this phenomenon, some have suggested that it is better to create thicker markets by not disclosing information, thus "bundling" many different impressions together (Ghosh, Nazerzadeh and Sundararajan 2007, Even-Dar, Kearns and Wortman 2007, McAfee, Papineni and Vassilvitskii 2010). The question of how to optimally bundle is a subject of ongoing

research (Bergemann, Bonatti and Said 2011).

Since targeting increases total surplus, platforms would like to allow targeting while still extracting the surplus this creates. This paper outlines a new and simple mechanism for doing so. We call it *buy-it-now or take-a-chance* (BIN-TAC), and it works as follows. Goods are auctioned with a buy-it-now price p, set relatively high. If a single bidder is willing to pay the price, they get the good for price p. If more than one bidder takes the buy-it-now option, a second price auction is held between those bidders with reserve p. Finally, if no-one participates in buy-it-now, an auction is held in which the top d bidders are eligible to receive the good, and it is randomly awarded to one of them at the $(d+1)$-st price.

In this manner, we combine the advantages of an auction and a fixed price mechanism. When matches occur, advertisers can self-select into the fixed-price buy-it-now option, allowing for revenue extraction. Advertisers are incentivized to do take the "buy-it-now" option because in the event that they "take-a-chance" on winning via auction, there is a significant probability they will not win the impression, even if their bid is the highest. On the other hand, when no matches occur, the auction mechanism ensures the impression is still allocated, thereby earning revenue.

BIN-TAC is simple, both in that it is easy to explain to advertisers and in that it requires relatively little input from the mechanism designer: a choice of buy-it-now price, randomization parameter d and optionally a reserve in the take-a-chance auction. As we show both analytically and through monte carlo simulation, BIN-TAC generally outperforms the two leading alternatives: a second price auction with reserve, or the "bundling" solution in which the platform withholds targeting information. At least in principle one could do better still by using the revenue-optimal mechanism suggested in Myerson (1981), which is considerably more complicated. We demonstrate that in our context BIN-TAC closely approximates the allocations and payments of the optimal mechanism, achieving similar performance.

To analyze its performance in a real-world setting, we turn to historical data from the Microsoft Advertising Exchange. By estimating the distribution of advertiser valuations, we can simulate the effect of introducing the BIN-TAC mechanism. We also consider a bundling strategy in which all impressions on a given webpage browsed by a user located in a particular geographic region are sold as identical products. We find that the optimal BIN-TAC mechanism generates 4.5% more revenue than the optimal second-price auction, while at the same time improving consumer surplus by 11%. This is possible because the optimal

second-price auction uses a high reserve to extract surplus from the long tail of valuations, whereas the BIN-TAC mechanism does this through a high buy-it-now price, which avoids excluding low valuation bidders. Both outperform the bundling strategy, although we cannot rule out better performance from an optimal bundling strategy.

We view the main contribution of our paper as introducing and analyzing a new and simple price discrimination mechanism that makes use of randomized auctions, and then testing its performance in a realistic environment. While our focus is on the display advertising market, we note that there are other markets in which randomized allocations are used as a screening tool. For example, Priceline offers users the choice between a hotel of their choice at a fixed high price, or the opportunity to bid for a random hotel room of certain guaranteed characteristics (e.g. location, star rating).

A secondary contribution of the paper is to document participation and bidding behavior in the display advertising market. While there has been theoretical work on this market (Muthukrishnan 2010, McAfee 2011), and some empirical work on the search advertising market (Ostrovsky and Schwarz (2009), Athey and Nekipelov (2010)), there has been little empirical work on display advertising. We document that there is a large gap between the highest and second highest valuations in these auctions, consistent with targeting creating thin markets. We also show that advertisers vary their bids based on the location of their users, taking advantage of user demographics provided by the platform to achieve better matches. Overall this work supports the assumptions typically made in the theoretical papers cited above.

Related Work: Our work is related to the literature on price discrimination and screening. Here we consider a mechanism that treats all bidders symmetrically, and proceeds sequentially. Other papers have suggested sequential screening approaches. Courty and Li (2000) consider a setting where the buyers themselves learn their type dynamically, in two stages. In this case, offering contracts after the first type revelation but before the second may be optimal; see Bergemann and Said (2010) for a survey on dynamic mechanisms. In the static setting, sequential screening and posted-price mechanisms can be used to design optimal (or near-optimal) mechanisms when the bidders have multi-dimensional private information (see for example Rochet and Chone (1998) and Chawla, Hartline, Malec and Sivan (2010)).

More generally, the question of whether sellers should provide information that allows buyers to target their bids arises in the analysis of optimal seller disclosure (see for example Lewis

and Sappington (1994) and Bergemann and Pesendorfer (2007)). The idea of bundling goods together to take advantage of negative correlation in valuations — in this case the negative correlation in the valuations from "match" or "no match" — dates back to Adams and Yellen (1976); see also McAfee, McMillan and Whinston (1989). Our paper is similar in style to Chu, Leslie and Sorensen (2011), who combine theory, simulations and empirics to argue that bundle-size pricing is a good approximation to the more complicated (but theoretically superior) mixed bundling pricing scheme for a monopolist selling multiple goods.

Finally, our model considers only the private values setting. Abraham, Athey, Babioff and Grubb (2010) consider an adverse selection problem that arises in a pure common value setting when some bidders are privately informed. This is motivated by the case when some advertisers are better able to utilize the user information provided by the platform. They show that asymmetry of information can sometimes lead to low revenue in this market.

From an empirical perspective, our paper contributes to the growing literature on online advertising and optimal pricing. Much of the work here is experimental in nature — for example, Lewis and Reiley (2011) ran a randomized experiment to test advertising effectiveness, while Ostrovsky and Schwarz (2009) used an experimental design to test the impact of reserve prices on revenues. There has also been recent work on privacy and targeting in online advertising (Goldfarb and Tucker 2011b, Goldfarb and Tucker 2011a).

Organization: The paper proceeds in three parts. First, we give an overview of the market for display advertising. In the second part we introduce a stylized environment, and prove existence and characterization results for the BIN-TAC mechanism. We also provide analytic results concerning the revenue maximizing parameter choices, and compare our mechanism to others using both theory and monte carlo simulation. Finally, in the third part we provide an empirical analysis of a display advertising marketplace, including counterfactual simulations of our mechanism's performance. All proofs are contained in the appendix.

2 The Display Advertising Market

This paper proposes a new second degree price discrimination strategy for advertising platforms such as Microsoft, Google and Facebook. In these markets, advertisers care about the characteristics of the users they advertise to, but it is up to the platform to choose whether

or not to disclose what they know about their users. The online display advertising market is an example of such a market. Its organization is depicted in Figure 1. On one side of the market are the "publishers": these are websites who have desirable content and therefore attract Internet users to browse their sites. These publishers earn revenue by selling advertising slots on these sites.

The other side of the market consists of advertisers. They would like to display their advertisements to users browsing the publisher's websites. They are buying user attention. Each instance of showing an advertisement to a user is called an "impression". Advertiser demand for each impression is determined by which user they are reaching, and what the user's current desires or intent are. For example, a Ferrari dealer might value high income users located close to the dealership. A mortgage company might value people that are reading an article on "how to refinance your mortgage" more than those who are reading an article on "ways to survive your midlife crisis", while the dealership might prefer the reverse.

Some large publishers, primarily AOL, Microsoft and Yahoo!, sell directly to advertisers. Since the number of users browsing such publishers is extremely large (e.g. 1.5% of total worldwide Internet pageviews are on Yahoo![1]), they can predict with high accuracy their user demographics. Consequently, they think of themselves of having a known inventory, consisting of a number of products in well-defined buckets: for example, male 15-24 year olds living in New York City viewing the Yahoo! homepage. They can thus contract to sell 1 million impressions delivered to a target demographic to a particular advertiser. Provided they have the inventory, they should be able to fulfill the contract. Transactions of this kind are generally negotiated between the publisher and the advertiser.

Alternatively, content is sold by auction through a centralized platform called an advertising exchange. Examples of leading advertising exchanges include the Microsoft Advertising Exchange (a subset of which we examine in this paper), Google's DoubleClick, and Yahoo's RightMedia.[2] Advertising exchanges are a minor technological wonder. They work in real-time. When a user loads a participating publisher's webpage, a "request-for-content" is sent to the advertising exchange. This request will specify the type and size of advertisement to be displayed on the page, as well as information about the webpage itself (potentially including information about its content), and information about the user browsing the page.[3]

[1] Source: alexa.com

[2] "In Sept 2009, RightMedia averaged 9 billion transactions a day with hundreds of thousands of buyers and sellers." Muthukrishnan (2010)

[3] For example, it may include their IP address and cookies that indicate their past browsing behavior.

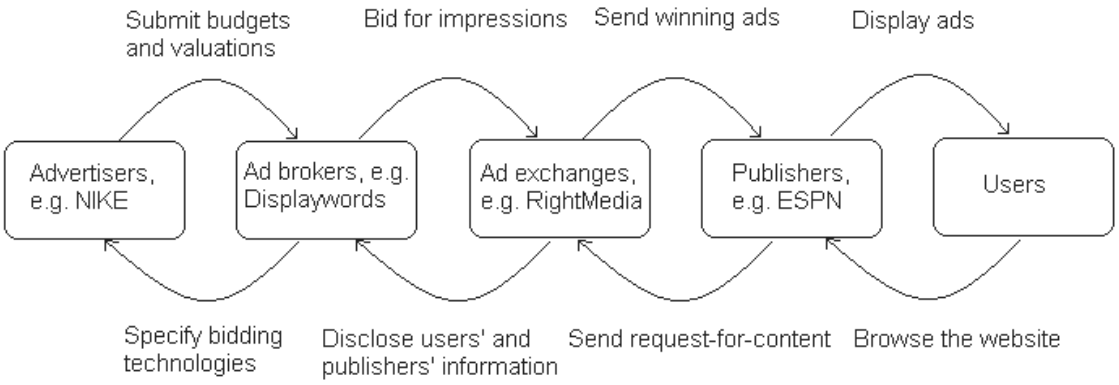


Figure 1: **The Display Advertising Market.**

The advertising exchange will then either allocate the impression to an advertiser at a previously negotiated price, or hold a second-price auction between participating advertisers. If an auction is held, all or some of the information about the webpage and user is passed along to ad brokers who bid on behalf of the advertisers. These ad brokers can be thought of as proprietary algorithms that take as input an advertiser's budget and preferences, and output decisions on whether to participate in an auction and how much to bid. The winning bidder's ad is then served by the ad exchange, and shown on the publisher's webpage.[4]

The bids placed in the auction are jointly determined by the preferences advertisers have, the ad broker interface and the disclosure policies of the ad exchanges or the publishers they represent. The ad brokers can only condition the bids they place on the information provided to them: if the user's past browsing history is not made available to them, they can't use it in determining their bid, even if their valuation would be influenced by this information. Similarly, the advertisers are constrained in expressing their preferences by the technology of the ad broker: if the algorithm doesn't allow the advertiser to specify a different willingness to pay based on some particular user characteristic, then this won't show up in their bids.

Ad exchanges have two main advantages over direct negotiation. First, they economize on transaction costs, by creating a centralized market for selling ad space. Second, they allow

[4]To make things yet more complicated, in some ad exchanges — though not Microsoft Advertising Exchange — two different pricing models coexist. The first is pay-per-impression, which is what we analyze in the current paper; the second is pay-per-click, where the payment depends on whether or not the user clicks on the advertisement. Ad exchanges use expected click through rates to compare these different bids through a single expected revenue number.

for very detailed products to be sold, such as the attention of a male 15-24 year old living in New York City viewing an article about hockey that has previously browsed articles about sports and theater. There is no technological reason why the products need to be sold in "buckets", as publishers tend to do when guaranteeing sales in advance. This "real-time" sales technology is often touted as the future of this industry, as it potentially improves the match between the advertiser and their target audience. We will focus on developing a real-time pricing mechanism for display advertising exchanges.

3 Model and Analysis

3.1 The Environment

A seller (publisher) has an impression to sell in real time, and they have information about the user viewing the webpage, summarized in a cookie. The seller is considering one of two policies: either disclosing the cookie content to the advertiser (the "targeting" policy), or withholding it (the "bundling" policy). When they allow targeting, bidders know whether the user is a "match" for them or not. When a match occurs, the bidder has a high valuation. But the probability of a match is low and matches are assumed independent, so it is likely that everyone in the auction has a low valuation. Allowing targeting may make the market "thin" in the sense of bids being relatively low.

Instead the seller may choose to withhold the cookie, so that bidders are uncertain about whether the user is a match for them or not. The seller thus bundles good impressions with bad ones, so that bidders have intermediate valuations. This reduces match surplus, but also reduces the bidder's information rents and so may be good for revenue.

The formal model is as follows. There are n symmetric bidders who participate in an auction for a single good which is valued at zero by the seller. Bidders are risk neutral. They have value V_H for the good when a match occurs, and value V_L for the good if no match occurs, where $V_L \sim F_L$ and $V_H \sim F_H$. We assume that F_L has support $[\underline{\omega}_L, \overline{\omega}_L]$ and F_H has support $[\underline{\omega}_H, \overline{\omega}_H]$, and that these supports are disjoint (so $\overline{\omega}_L < \underline{\omega}_H$). We assume both F_L and F_H have continuous densities f_L and f_H. The Bernoulli random variable X indicates whether a match has occurred, and the event $X = 1$ occurs with probability $\alpha \in (0, 1)$.

The bidder type is a triple (X, V_L, V_H) is drawn identically and independently across bidders,

so that a user who is a match for one advertiser need not be a match for the others. In the case with targeting, each advertiser's realized valuation $V = (1 - X)V_L + XV_H$ is private information, known only to the advertiser. Instead if the seller bundles all impressions, the advertiser knows V_L and V_H but does not know the realization of X, implying their expected valuation is $E[V] = (1 - \alpha)V_L + \alpha V_H$.

For simplicity of the presentation, we also make some technical assumptions on the *virtual valuations* $\psi(v) = v - \frac{1-F(v)}{f(v)}$.[5] We assume that $\psi(v)$ is continuous and increasing over the regions $[\underline{\omega}_L, \overline{\omega}_L]$ and $[\underline{\omega}_H, \overline{\omega}_H]$. We additionally assume that $\psi(v)$ single-crosses zero, that this intersection occurs in the low valuation region $[\underline{\omega}_L, \overline{\omega}_L]$, and that $\psi(\overline{\omega}_L) \leq \psi(\underline{\omega}_H)$. Overall, our environment is fully characterized by the tuple (n, α, F_L, F_H).

Discussion: We assume that the match random variables X and the valuations V_L and V_H are independent across bidders. We focus on independence for two reasons. First, it is an assumption that is often made in the screening and mechanism design literatures, and so is a natural starting point. Second, in the log data examined in this paper we observe little correlation in bids.[6]

We also will focus on environments where α is small, since this implies that the probability of zero or a single match is high. This is the interesting case, reflecting the industry concern that providing "too much" targeting information reduces competition and hurts revenues. In our data we often observe a large gap between the highest and second highest bid, which provides support for this focus.

3.2 Pricing Mechanisms

We propose using a randomized auction as a pricing mechanism. Our BIN-TAC mechanism works as follows. A *buy-it-now price* p is posted. Buyers simultaneously indicate whether they wish to *buy-it-now* (*BIN*). In the event that exactly one bidder elects to buy-it-now, that bidder wins the auction and pays p. If two or more bidders elect to BIN, a second-price sealed bid auction with reserve p is held between those bidders. Bidders who chose to

[5] Without these assumptions we would have to analyze multiple cases, which is straightforward but tedious.

[6] In practice, the information that platforms may choose to disclose is multidimensional, and some user characteristics may be "vertical" (e.g. income) and therefore induce positive correlation in match probabilities; while others may be "horizontal" and have correlation implications that depend on the population of advertisers (e.g. age).

BIN are obliged to participate in this auction. Finally, if no-one elects to BIN, a sealed bid *take-a-chance* (TAC) auction is held between all bidders, with a reserve r. In that auction, one of the top d bidders is chosen uniformly at random, and if that bidder's bid exceeds the reserve, they win the auction and pay the maximum of the reserve and the $(d+1)$-th bid. Ties among d-th highest bidders are broken randomly prior to the random allocation. We call r the TAC-reserve, and d the randomization parameter.

To analyze the performance of BIN-TAC, it will be useful to have some benchmarks for comparison. A natural benchmark is the pricing mechanism that is most commonly used in practice, the second price auction (SPA). We distinguish between when an SPA is used and targeting is allowed (SPA-T), and when it is used with bundling (SPA-B).

A third benchmark is the revenue-optimal mechanism within the class of those that allow targeting (i.e. those that commit to reveal the cookie to all bidders for free).[7] Usually this mechanism is the second-price auction with an optimally chosen reserve price. However in this case the virtual valuations $\psi(v)$ are not increasing over the whole support of F — indeed they are (infinitely) negative over the region $(\overline{\omega}_L, \underline{\omega}_H)$. The optimal mechanism may require ironing (Myerson 1981).

In plain terms, ironing implies that sometimes the allocation will be randomized among bidders with different valuations. Just as in our TAC auction, the winner of the auction need not have the highest valuation. The difference is that in the optimal mechanism, the randomization only takes place when two or more bidders — including the highest valuation bidder — have valuations in a given "ironing" region. By contrast, in BIN-TAC this randomization occurs whenever no-one takes the BIN option. The differences will be clearer later when we compare the performance of the mechanisms. For now, we would like to present a simple example to illustrate why our BIN-TAC mechanism may be good at delivering both social surplus and revenue, as a motivation for the detailed equilibrium analysis that follows.

3.3 A Motivating Example

Consider a special case of our environment with just two bidders, and a match probability of 10%. Bidders have fixed symmetric valuations, equal to 10 if they match, and 1 otherwise.

[7] A seller may potentially do better by withholding match information altogether (bundling), or by selling the rights to the match information — see Bergemann and Pesendorfer (2007).

Now consider the expected outcomes of the two second-price auction mechanisms. With targeting, the allocation will be fully efficient. The probability that at least one bidder has a high valuation is $1-(0.9)^2 = 0.19$, and so expected surplus is $10(0.19) + 1(0.81) = 2.71$. On the other hand, the probability that both bidders match is only 1%, so expected revenue is only $(0.01)10 + (0.99)1 = 1.09$.

Under bundling, the two bidders don't know if the impression is a match, and so value it at its expected value of $(0.1)10 + (0.9)1 = 1.9$. They bid identically, yielding expected revenues of 1.9. This is a significant improvement. But now the allocation may be ex-post inefficient, with the lower valuation bidder getting the impression. Expected surplus is equal to $(0.01)10 + (0.18)5.5 + (0.81)1 = 1.9$, much lower than before. Notice that the bundling strategy has eliminated all the buyers' information rents, so that the seller captures all the surplus as revenue.

Next, consider the BIN-TAC mechanism with a BIN price of 5.5, TAC reserve of 1, and randomization parameter 2. When a buyer matches, they will (weakly) take the BIN price, since their surplus on doing so is $10 - 5.5 = 4.5$, whereas if they take-a-chance, they have a 50% chance of getting the object, with expected payoff $(0.5)(10-1) = 4.5$. So if both buyers match, there will be an auction with revenue 10; if one buyer matches, the revenue will be 5.5; and if none match it will be 1. Adding this up gets an expected revenue of 1.9, as in the SPA-B case. But notice that the BIN-TAC allocation is fully efficient, and thus gets the same surplus as the SPA-T.[8] So the BIN-TAC mechanism may improve revenues relative to the SPA-T, and welfare relative to the SPA-B.

3.4 Equilibrium Analysis

Moving back to the general environment, we characterize equilibrium strategies under BIN-TAC. We proceed by backward induction. The auctions that follow the initial BIN decision admit simple strategies. If multiple players choose to BIN, the allocation mechanism reduces to a second-price auction with reserve p. Thus, it is weakly dominant for players to bid their valuations.[9]

Truth-telling is also weakly dominant in the TAC auction. The logic is standard: if a bidder

[8] It is revenue-optimal within the class of mechanisms with targeting, so we needn't analyze that separately.

[9] Since participation is obligatory at this stage, the minimum allowable bid is p; but no bidder would take the BIN option unless they had a valuation of at least p.

with valuation v bids $b' > v$, it can only change the allocation when the maximum of the d-th highest rival bid and the reserve price is in $[v, b']$. But whenever this occurs, the resulting price of the object is above the bidder's valuation and if she wins she will regret her decision. Alternatively, if she bids $b' < v$, when she wins the price is not affected, and her probability of winning will decrease.

Taking these strategies as given, we turn to the buy-it-now decision. Intuitively, the BIN option should be more attractive to higher types: they have the most to lose from either random allocation (they may not get the good even if they are willing to pay the most) or from rivals taking the BIN option (they certainly do not get the good). This suggests that in a symmetric equilibrium, the BIN decision takes a threshold form: $\exists\ \overline{v}$ such that types with $v \geq \overline{v}$ elect to BIN, and the rest do not. This is in fact the case.

Prior to stating a formal theorem, we introduce the following notation. Let the random variable Y^j be the j-th highest draw in an iid sample of size $n-1$ from F (i.e. the j-th highest rival valuation) and let Y^* be the maximum of Y^d and the TAC reserve r.

Theorem 1 (Equilibrium Characterization)
Assume $d > 1$ and $p \leq \frac{d-1}{d}\overline{\omega}_H + \frac{1}{d}E[Y^*]$. Then there exists a unique symmetric pure strategy Bayes-Nash equilibrium of the game, characterized by a threshold $\overline{v}$ satisfying:

$$\overline{v} = p + \frac{1}{d}E\left[\overline{v} - Y^* | Y^1 < \overline{v}\right] \tag{1}$$

Types with $v \geq \overline{v}$ take the BIN option; and all types bid their valuation in any auction that may occur.

Equation (1) is intuitive: Which type is indifferent between the BIN and TAC options? If strategies are increasing, the only time the choice is relevant is when there are no higher valuation bidders (since otherwise those bidders would BIN and win the resulting auction). So if a bidder has the highest value and chooses to BIN, they get a surplus of $v - p$. Choosing to TAC gives $\frac{1}{d}E\left[v - Y^* | Y^1 < v\right]$, since they only win with probability $\frac{1}{d}$, although their payment of Y^* is on average much lower. Equating these two to find the indifferent type $\overline{v}$ yields Equation (1).[10]

[10]The assumption that $p \leq \frac{d-1}{d}\overline{\omega}_H + \frac{1}{d}E[Y^*]$ rules out uninteresting cases where the BIN price is so high that no-one ever chooses BIN.

Now we consider the revenue-maximizing choices of the design parameters: the BIN price p, the TAC reserve r and the randomization parameter d. It is hard to characterize the optimal d, as it is an integer programming problem which doesn't admit standard optimization approaches. However for a given d, the optimal BIN price and TAC reserve are given by some familiar looking equations. Again, we must introduce some notation. Let $p(\overline{v}, r) = \overline{v} - \frac{1}{d} E\left[\overline{v} - Y^* | Y^1 < \overline{v}\right]$ be the solution of Equation (1), expressing the BIN price as a function of the threshold and the TAC reserve. Let $R(\overline{v}, r)$ be the conditional expected revenue from a TAC auction when the highest valuation is exactly equal to $\overline{v}$ and the reserve is r. Then we have the following theorem.

Theorem 2 (Optimal Buy Price and Reserve) *For any (p, d), the revenue-maximizing TAC reserve r^* satisfies:*

$$r^* = \frac{1 - F(r^*)}{f(r^*)} \tag{2}$$

The optimal BIN price is given by $p(\overline{v}^*, r^*)$ where $\overline{v}^*$ is the solution of the equation below:

$$F(\overline{v})\left(p(\overline{v}, r^*) - R(\overline{v}, r^*)\right) + (n-1)(1 - F(\overline{v})\left(\overline{v} - p(\overline{v}, r^*)\right) = \frac{(1 - F(\overline{v}))F(\overline{v})}{f(\overline{v})} \frac{\partial p(\overline{v}, r^*)}{\partial \overline{v}} \tag{3}$$

If no such solution exists in $[\underline{\omega}_H, \overline{\omega}_H]$, then the optimal BIN price is $p(\underline{\omega}_H, r^*)$.

Equation (2) is somewhat surprising; the optimal TAC reserve is exactly the standard reserve in Myerson (1981), ensuring that no types with negative virtual valuation are ever awarded the object. This is despite the fact that our BIN-TAC mechanism is not the optimal mechanism. The key insight is that the TAC reserve is relevant for the BIN choice. Raising the TAC reserve lowers the surplus from participating in the TAC auction, and so the seller can also raise the BIN price while keeping the indifferent type $\overline{v}$ constant. So the trade-off is exactly the usual one: raising the TAC reserve extracts revenue from types above r^* — even those above $\overline{v}$ — at the cost of losing revenue from the marginal type. This is why we get the usual solution.

On the other hand, the implicit equation for the optimal BIN price is new. Notice that the BIN price in some sense sets a reserve at $\overline{v}$. If two bidders meet the reserve, the seller gets the second highest bid; if only one, the BIN price; and if none, he gets the TAC revenue. So a marginal increase in the threshold has three effects. First, if the highest bidder has valuation exactly equal to the threshold, following an increase she will shift from BIN to TAC. This costs the seller $p(\overline{v}, r^*) - R(\overline{v}, r^*)$. Second, if the second highest bidder has valuation equal

to the threshold, an increase will knock her out of the BIN auction, and the seller's revenue falls by $\overline{v} - p(\overline{v}, r^*)$. Finally, if the highest bidder is above the reserve and the second highest is below, an increase gains the seller $\frac{\partial p(\overline{v}, r^*)}{\partial \overline{v}}$. Working out the probabilities of these various events, and equating expected costs and benefits, we get the result.

Sometimes there is no solution for $\overline{v}^*$ in $[\underline{\omega}_H, \overline{\omega}_H]$. This occurs whenever the high valuations are substantially larger than the low valuations (i.e. $\underline{\omega}_H \gg \overline{\omega}_L$), so that it is not profitable to randomize the allocation for any of the high types. In this case the BIN price is set at $p(\underline{\omega}_H, r^*)$ so that the lowest high type at $\underline{\omega}_H$ elects to BIN.

3.5 Performance Comparisons

We are interested in comparing the BIN-TAC mechanism to the benchmarks in terms of both revenue and total welfare. For any mechanism M with parameters θ, define a payoff function $\pi(M, \theta, \beta)$ as follows (suppressing the dependence on the environment):

$$\pi(M, \theta, \beta) = ER(M, \theta) + \beta ECS(M, \theta) \tag{4}$$

where ER denotes expected revenue and ECS expected consumer surplus. Notice that when $\beta = 0$, the platform objective is just to maximize revenue as in the usual optimal mechanism design problem. Similarly, when $\beta = 1$ the objective aligns with the social planner problem of maximizing welfare. We say that mechanism M dominates M' over the interval $[a, b] \subseteq [0, 1]$ if $\max_\theta \pi(M, \theta, \beta) \geq \max_\theta \pi(M', \theta, \beta)$ for all $\beta \in [a, b]$ and for all environments (n, α, F_L, F_H). If M dominates M' over the whole interval $[0, 1]$ we say that M dominates M'. We say such dominance is strict if for some environment and some β the inequality holds strictly. Strict dominance means that regardless of whether the platform is maximizing revenue, joint welfare, or some combination of the two, and regardless of the environment, mechanism M is better able to achieve that objective than M'.

Theorem 3 (Mechanism Performance) *(i) BIN-TAC strictly dominates SPA-T. (ii) For any environment* $\exists\ \underline{\beta} < 1$ *such that BIN-TAC dominates SPA-B on* $(\underline{\beta}, 1]$. *(iii) In the special case when* F_H *and* F_L *are degenerate with atoms at* V_H *and* V_L *respectively, BIN-TAC strictly dominates SPA-B.*

The formal proof is in the appendix, but we provide some intuition here. The first result follows by showing that SPA-T is just a special case of BIN-TAC, and therefore any performance achievable by SPA-T is also achievable by BIN-TAC. The idea is to turn the TAC auction into an SPA, by setting the randomization parameter $d = 1$ and the BIN-price p so high that the BIN option is never taken. The second result follows by showing that BIN-TAC is always better at achieving an efficient allocation (since the bundling solution suppresses the information needed to ensure good match outcomes), and therefore as long as the weight on consumer surplus is sufficiently high, there is some interval of weights for which BIN-TAC is better.

The next question is whether BIN-TAC could be a good solution even when the platform is only interested in revenue maximization. Our last statement shows that this is possible: when the only source of private information is the match variable X, disclosing that information and then using BIN-TAC is better than withholding it and running a second-price auction. This is interesting, as it is natural to assume that bundling is most effective when it removes the only source of private information, therefore eliminating all information rents. But in this case, running a TAC auction causes no distortion in the allocations at the bottom (a TAC auction is run only when all bidders have valuations V_L), and so BIN-TAC is able to extract most of the information rents it creates and do better.

By definition, BIN-TAC will have (weakly) worse revenue performance than the revenue-optimal mechanism. The question then is how close BIN-TAC gets. We will show, informally, that it gets very close indeed, using graphs and simulations. At this point it will be useful to describe the revenue-optimal mechanism in detail. Because it is hard to solve for analytically, to economize on space we derive it in the supplementary appendix. There we also show that the interesting case occurs when $\alpha\underline{\omega}_H < r^*(1 - F(r^*))$, where r^* is the optimal reserve of equation (2).[11] In that case, define the ironed virtual valuations as follows:

$$\phi(v) \quad = \quad \begin{cases} 0 & v \in [\underline{\omega}_L, r^*) \\ \psi(v) & v \in [r^*, \widetilde{v}] \\ \psi(\widetilde{v}) & v \in (\widetilde{v}, \underline{\omega}_H) \\ \psi(v) & v \in [\underline{\omega}_H, \overline{\omega}_H], \end{cases} \tag{5}$$

[11] When this condition fails, ironing is not required and the optimal mechanism is just a second price auction with reserve, which can be implemented as a BIN-TAC auction with $d = 1$.

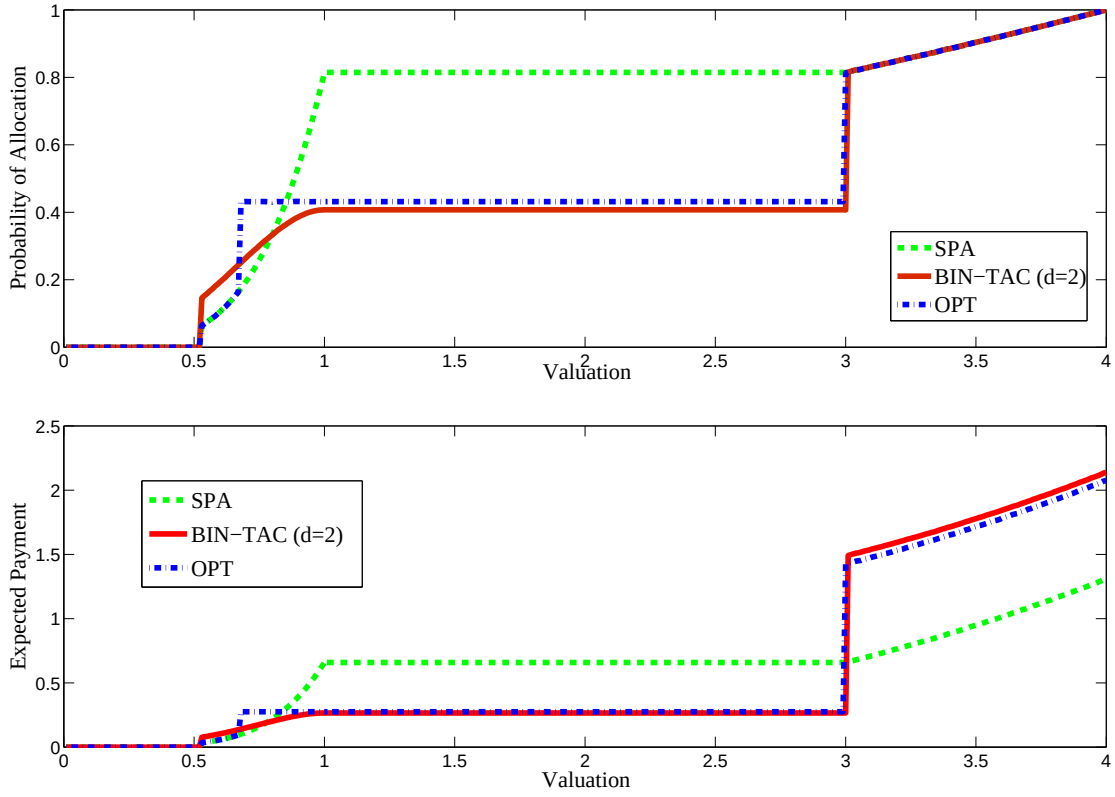


Figure 2: **Comparison of Allocations and Payments.** Allocation probabilities (top panel) and expected payments (bottom panel) for the OPT, SPA and BIN-TAC mechanisms when the distributions F_L and F_H are uniform. The x-axis corresponds to the bid.

where $r^* \leq \widetilde{v}\underline{\omega}_L$.The allocation procedure works like this: award the good to the bidder with the highest ironed virtual valuation, breaking ties at random, provided the virtual valuation is positive. Notice that all types between $\widetilde{v}$ and $\underline{\omega}_H$ get the same ironed virtual valuations, and therefore if they tie, the winner is selected at random. Like BIN-TAC, this is inefficient, but allows additional revenue extraction from higher types.

Having obtained this characterization, we can compare BIN-TAC with the optimal mechanism. For now, let us focus on a simple environment, where F_L is uniform over $[0, 1]$ and F_H is uniform over $[\Delta, \Delta+1]$. Figure 2 shows the interim allocation probabilities (top panel) and expected payments by type (bottom panel) as a function of bidder type, in the case where $\Delta = 3$, $\alpha = 0.05$ and $n = 5$ (with optimal parameter choices). The optimal mechanism has a discontinuous jump in the allocation probability at $\widetilde{v} = 0.676$, and then irons until the high

valuation region on $[3, 4]$. As you can see, BIN-TAC is able to approximate the discontinuous increase in allocation probability at $\widetilde{v}$ with a smooth curve, by randomizing the allocation in that region using the TAC auction. By contrast, the slope of the SPA-T allocation schedule is steep on this region and so the SPA cannot extract revenue from the high types (who could easily pretend to be a lower type while barely changing their probability of winning). This is clear from the bottom panel.

Table 1 compares the expected revenue and welfare obtained by all the mechanisms. The performance of BIN-TAC is close to the optimal mechanism (about 96% of OPT), much better than the optimal SPA-T (85%). The table also shows that SPT-B performs less well than both BIN-TAC and OPT, especially in terms of expected consumer surplus. This is because it often fails to match advertisers and users correctly.

Table 1: Revenue Comparison: Uniform Environment

Mechanism	OPT	SPA-T	BIN-TAC (d=2)	BIN-TAC (d=3)	SPA-B
Expected Revenue	0.89	0.76	0.85	0.83	0.81
Expected Consumer Surplus	0.51	0.67	0.48	0.40	0.16

Expected revenue and welfare under different mechanisms, for the uniform environment with $\Delta = 3$, $\alpha = 0.05$ and the number of bidders $n = 5$.

3.6 Monte Carlo Simulations

We would like to test our mechanism against the benchmarks in a variety of other settings. We drop the assumption that F_L and F_H have disjoint support. The optimal BIN-TAC mechanism remains easy to calculate. Nothing in the proof of Theorem 2 required the disjoint supports for determining r^* and p^*, and so these can be solved for numerically for each d. Thus the optimization problem reduces to a one dimensional discrete optimization problem in the randomization parameter d, which can be quickly solved. Finding the optimal mechanism is more challenging, but can be done using standard optimization techniques.

For our simulations, we restrict ourselves to location families where the distribution $F_H(\cdot) = F_L(\cdot - \Delta)$ for some shift-parameter Δ, as in the uniform case above. This Δ is the difference in mean valuation between the high and low groups, which we call the "match increment". We consider two location families: one where V_L is normal, and another where V_L is log

normally distributed. In both cases V_L has mean 1 and standard deviation 0.5. We allow Δ, n and α to vary across experiments, and compute r^*, p^* and d^* as discussed. The default parameters we consider are $n = 10$, $\Delta = 5$, and $\alpha = .05$, and we vary one parameter at a time. Each experiment is repeated for 100000 impressions, and we calculate the average revenues.

The results are presented in Figures 5, 6 and 7. In all cases, on the y-axis we plot the revenue as a fraction of the revenue from the optimal mechanism. Recall that BIN-TAC generalizes SPA-T, so its performance is always at least as good, and often significantly better. In all cases, the BIN-TAC extracts at least 90% of the optimal revenue, compared to a worst-case performance of around 82% for the SPA-T. Consistent with Theorem 3, the SPA-B in some cases does even better than OPT (when there are very few bidders), but its performance sharply degrades as the probability or value of a match gets large.

We see this in Figures 5 and 6. The expected number of matches is αn, and so as either α or n increases, the performance of the mechanisms that allow targeting improves relative to the SPA-B. Over some range, BIN-TAC also significantly outperforms the SPA-T, but as the number of bidders or the probability of match get sufficiently high, both converge to the OPT mechanism (which is itself an SPA with high reserve).

Figure 7 shows the dependence on the gap Δ. As expected, the performance of BIN-TAC increases while that of SPA-T falls as Δ gets larger, over some range. Since there is more revenue to be gained from high-valued bidders, BIN-TAC can only perform better with a large Δ. For sufficiently high Δ though, both BIN-TAC and the SPA set high reserves, "throwing away" low-valued impressions and extracting all their revenue from matches, with equal revenue performance.

Overall, the performance of BIN-TAC is very good, at least for the distributions and parameters chosen. The main caveats are that it doesn't perform well with very few bidders (when bundling is preferable), and has little to recommend it when matches are highly probable or very valuable (a second-price auction would do as well). Its niche is in markets with relatively large numbers of bidders but low match probabilities, so that markets are "thin" in the sense of having relatively low matches in expectation.

4 Empirical Application

Our theoretical analysis has shown that there are cases in which BIN-TAC performs well. We now test our mechanism's performance in a real-world setting. We have historical data from Microsoft Advertising Exchange, one of the world's leading ad exchanges. Our data comes from a single large publisher's auctions on this exchange and consists of a 0.1% random sample of a week's worth of auction data from this publisher, sampled within the last two years. This publisher sells multiple "products", where a product is a URL-ad size combination (e.g. a large banner ad on the sports landing page of the New York Times). This data includes information on both the publisher and advertiser side. On the publisher side, we see the url of the webpage the ad will be posted on, the size of the advertising space and the IP address of the user browsing the website. We form a unique identifier for the url-size pair, and call that a product. We determine which US state the user IP originates from, and call that a region. We use controls for product and region throughout the descriptive regressions. Unfortunately, we don't have more detailed information on the product or the user, as the tags and cookies passed by the publisher to the ad exchange were not stored.

On the advertiser side, we see the bid they placed, the company name, the ad broker they employed, and a variable indicating the ad they intend to show. In the overwhelming majority of cases there is a single ad for each company, but some larger firms have multiple ad campaigns simultaneously. We treat these as being a single ad campaign in what follows because each firm should have the same per impression valuation across campaigns. We observe who won the auction and the final price.

We drop auctions in which the eventual allocation was determined by biased bids and modifiers.[12] We also restrict attention to impressions that originate in the US, and where the publisher content is in English. Finally, we restrict only to reasonably frequently sold products, those with at least 100 sales in the dataset. This leaves us with a sample of 83515 impressions.

The dataset is summarized in Table 2. For confidentiality reasons, bids have been rescaled so that the average bid across all observations is equal to 1 unit. Bids are very skew, with the median bid being only 0.57 units. Perhaps as a consequence of this skewness, the winning bid

[12]When the advertiser has a technologically complex kind of ad to display, their bid is modified. When the advertiser has a previously negotiated contract with the platform, their bid may be biased. This can affect the allocation and payments.

— which is more heavily sampled from the right tail of the bid distribution — is much higher at 2.96 units. There are on average 6 bidders per auction, but there is considerable variation in participation, with a standard deviation of nearly 3. Bids are not strongly correlated: as the table shows, the correlation between a randomly selected pair of bids from each auction is only 0.01. This is not statistically significant at 5% (p-value 0.116, $N = 15827$).[13]

The advertisers are themselves quite active in the market. On average they bid on 0.7% of all impressions, and win nearly 40% of those they bid on. These averages are somewhat misleading though. The median advertiser is far less active, bidding on only 0.02% of impressions, while the most active advertiser participates in nearly 90% of auctions. Our hypothesis is that some advertisers choose to participate in relatively few auctions, but tend to bid quite highly and therefore win with relatively high probability. Others bid lower amounts in many auctions, and win with lower probability. The first strategy is followed by companies who want to place their advertisements only on webpages with specific content or to target specific demographics, while the latter strategy is followed by companies whose main aim is brand visibility.

4.1 Descriptive Evidence

Before proceeding to the main estimation and simulations, we provide some evidence that advertisers bid differently on different users (i.e. there is matching on user demographics). We also show that the platform is doing poorly in extracting this match surplus as revenue.

Leading advertisers do vary their bids on the same product over short periods of time. Figure 3 shows re-scaled bids in 50 auctions by five large advertisers for the most popular webpage slot sold by this publisher. The advertisers were chosen at random from the top 50 advertisers in our dataset (ranked by purchases). The 50 auctions are chosen to be consecutive for each bidder.[14] The bids exhibit considerable variation, even though all of these impressions were auctioned within a 3-hour period. While this could in principle be driven by decreases in the advertisers' available budget, since the bids go both up and down it seems more likely that this variation arises from matching on user demographics.

[13]That bids are not positively correlated should not be taken to mean that underlying valuations are not positively correlated; it could just be that informational and technological constraints prevent advertisers from fully expressing their preferences.

[14]Since the same set of bidders don't participate in every auction, the impression number on the x-axis corresponds to different impressions for different bidders.

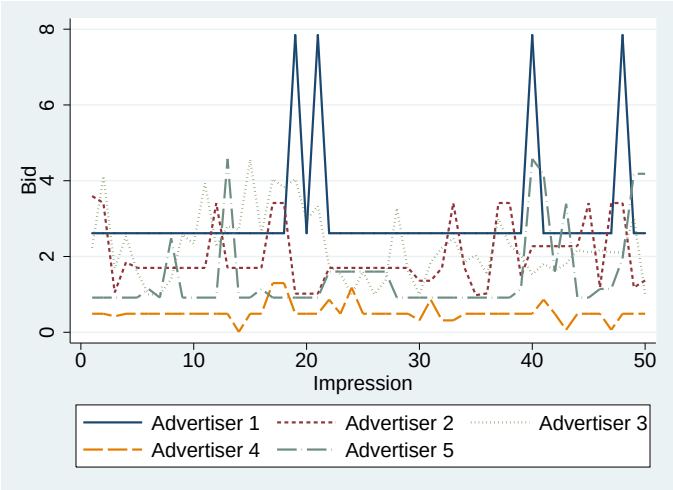


Figure 3: **Bids over Time.** The figure shows the (rescaled) bids of five advertisers in our data, selected at random from the top 50 advertisers (ranked by purchases) on 50 randomly chosen successive impressions of the most popular product. Note that the set of impressions differs across bidders (there are no impressions on which all 5 participate).

One direct test of advertiser-user matching is to look for the significance of advertiser-user fixed effects in explaining bids. Specifically, we estimate an unrestricted model where the dependent variable is bids and the controls are advertiser-user dummies, versus a restricted model with just advertiser and user fixed effects, but not their interaction. The restricted model is overwhelmingly rejected by the data: the relevant F-statistic is over 15, while the 99% critical value is just over 1. This points towards matching on demographics.

Proving that this matching is motivated by economic considerations is a little more difficult. The only user demographic we observe is the user region, and it is hard to know a priori what the advertisers' preferences over regions are. To get a handle on this, we turn to another proprietary dataset that indicates how often an advertiser's webpage was viewed by internet users in different regions of the country during the calendar month prior to the auction.[15] Our intention is to proxy for the advertisers' geographic preferences (insofar as these exist) using this pageview data. The idea is that firms who operate in only a few regions probably attract all their pageviews from those regions, and also only want to advertise in those regions. If this is right, advertisers who attract a large fraction of their pageviews from a particular region should participate more frequently and bid higher on users from those regions.[16] We normalize the pageviews from a particular state by the state population to

[15]For example, if these auctions were in May, the pageview data would be taken from April.

[16]Because the pageview data dates from a period before our exchange data we are not worried about

get a per capita pageview measure, and then construct the fraction of normalized pageviews each region receives, calling this the "pageview ratio".

In Table 3, we present results from regressions of auction participation (a dummy equal to one if the advertiser participated), and bid (conditional on participation) on the pageview ratio, as well as a number of fixed effects. Because the sheer size of our dataset makes it difficult to run the fixed effect regressions, we run this on a subsample consisting of the top 10% of advertisers.[17] The first column shows participation as a function of the pageview ratio, as well as product-region fixed effects, and time-of-day fixed effects (since participation and bids may vary with the user's local time). We find a positive but insignificant effect. But when we include advertiser fixed effects to control for different participation frequencies across advertisers, we find a much bigger and now highly significant effect. All else equal, an advertiser is 3.3% more likely to bid on a user from a state that contributes 10% of the population-weighted pageviews for their site. This is a large increase, as the average probability of participation is only around 1%.

Turning to the bids, we find similar estimates and significance levels from the specifications with and without advertiser fixed effects. We find that firms bid higher on users from more relevant regions, although this effect is relatively modest in economic terms. Given that our proxy for advertiser preferences is relatively crude, it is notable that we find these effects. This provides some evidence that the matching is surplus increasing, in that advertisers are able to target regions where their most valuable customers are.

A second stylized fact is that there is often a substantial gap between the highest and second highest bid in the auction. To facilitate bid comparisons, we look at the product with the highest sales volume in the data (over 38% of all impressions). The left panel of Figure 4 shows a kernel density estimate of this gap. The average bid in an auction is 0.88, while the mean gap is much larger at 1.89, indicating that there is a lot of money left on the table by a second-price mechanism (see Table 2 for other summary statistics). That gap itself is extremely skewed.

Assuming bids are equal to valuations — an assumption we will motivate in the next section — the right panel shows the virtual valuations $\psi(v)$ as a function of the bids. Although the

reverse causality (i.e. advertisers who win more impressions from region X later get more views from region X).

[17] Fortunately since participation is highly skewed, these advertisers account for 90% of the bids. With only bidder fixed effects we could use a within transformation to reduce the computational burden; but unfortunately this is not possible with multiple non-interacting fixed effects.

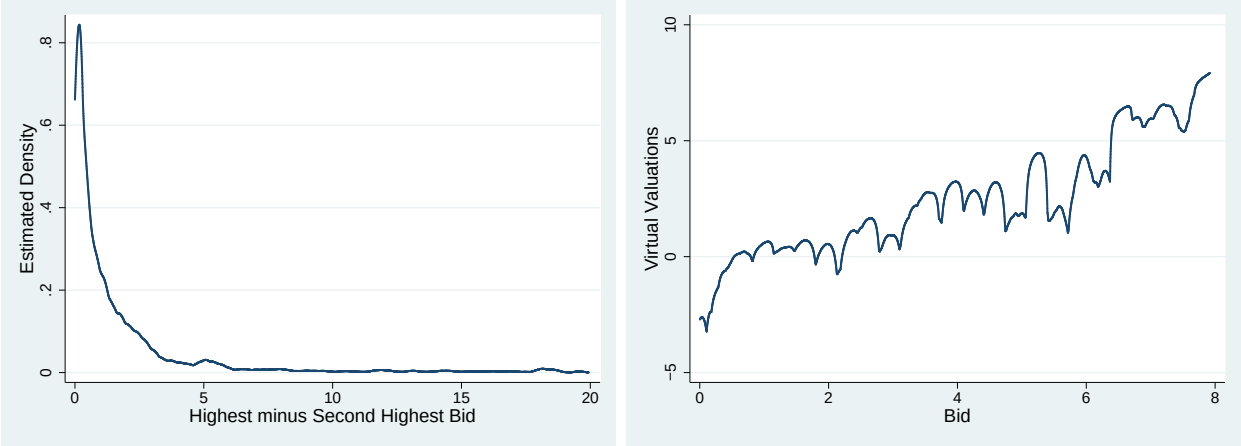


Figure 4: **Bidding Gap and Virtual Valuations.** The left panel shows a kernel density estimate of the pdf of the (normalized) gap between the highest and second highest (rescaled) bids in auctions for the product with the highest sales volume in our dataset. The right panel shows the estimated virtual valuations as a function of bids.

virtual valuations are never infinitely negative, as in our stylized model, they are certainly non-monotone. This implies that BIN-TAC may be able to extract more revenue than a second price auction. We test this in the next section.

4.2 Estimation and Counterfactual Simulations

Our theoretical model is of a single auction with a particular valuation structure, rather than a whole market with a general valuation structure, and so in order to provide micro-foundations for our simulation approach, we need to enrich the model.

We make the following assumptions for the estimation and counterfactual simulations. There is a fixed set of N bidders who are always present in the market.[18] As in the text, the model is symmetric independent private values. Each bidder draws their valuations for each impression identically, independently and privately according to some distribution F_j supported on $[0, \infty)$ (where j indexes products). So bidder valuations are independent both across bidders and within a bidder over time. This is a strong assumption, as it rules out common preferences for particular user demographics. For example, it rules out the

[18]The assumption that bidders are continuously present in the market is in principle relatively innocuous since bidding is done by ad broker algorithms. Yet some bidding algorithms ignore certain auctions in order to respect advertiser budget constraints. We will not model this "inattention", especially because it is hard to rationalize such behavior as optimal: bidding close to zero has almost no effect on the budget constraint since the maximum possible payment in a second-price auction is bounded above by the bid.

possibility that all bidders prefer high income bidders, in which case we would observe positive correlation in bids. Some partial support for this assumption comes from the lack of bid correlation reported in Table 2. The symmetry assumption is also strong — and probably rejected by the data given the significance of the advertiser fixed effects in the reduced form regressions — but helps to keep the problem computationally tractable. To address the concern that the symmetry and independence assumptions are driving our results, we will do some robustness tests based on different informational assumptions in a later subsection.

From the summary statistics we also know that participation varies across advertisers. We assume that participation costs are zero, and thus we can infer from non-participation that an advertiser has zero valuation for the impression (since with any positive valuation there is weakly positive surplus from bidding). This may seem like a strong assumption, but given that the 5th percentile of bids in our data is equal to 0.013 — tiny in real terms, with an almost zero probability of winning, and even lower surplus — it is hard to believe that participation costs are substantially different from zero. One reason for this may be that bidding is automated.

Given these assumptions, we are able to make the following inference from the second-price auction data. Letting $i = 1 \dots I$ index bidders and $t = 1 \dots T$ index auctions, if bidder i makes a bid of $b_{i,t}$ in auction t, their valuation is $b_{i,t}$, since it is weakly dominant for them to bid their valuations. Moreover, if bidder i did not participate in auction t, their valuation for that particular impression must have been zero. Since there is a one-to-one mapping from the distribution of bids and participation to the valuations, F_j is non-parametrically identified. We could therefore estimate the valuation density for each product using non-parametric methods. But, as we will show below, the counterfactual simulations will never require estimating more than some conditional moments of order statistics (e.g. the expected value of the d-th highest valuation when the highest valuation is less than $\overline{v}$). So instead we estimate these moments by the corresponding sample average.

We are interested in comparing the "optimal" BIN-TAC mechanism to other leading mechanisms. For simplicity, we restrict attention throughout to the class of mechanisms that make the same parameter choices for all products (e.g. we rule out different reserves or randomization parameters by product or user-region). In each case we find these optimal parameters by maximizing the revenue functions defined in equations (6) and (7) below, using standard optimization methods.[19] To get standard errors on our revenue and consumer

[19]This raises an over-fitting concern, in that the parameters are optimized for this specific realization of

surplus estimates, we bootstrap the estimation sample and re-run the simulation procedure, holding the parameter choices fixed.[20]

Mechanisms with Targeting: The two policies we want to compare here are the second price auction with targeting and BIN-TAC. The two SPA mechanisms are easiest. For example, with a reserve of r, the expected revenue depends on the joint distribution of the top two valuations: since bidders bid their valuations, the item sells if the highest valuation exceeds r, and then the revenue is the maximum of the second highest bid and r. Letting the k-th highest bid in an auction t be $b_t^{(k)}$, our estimate is then given by the sample average across the T auctions:

$$\text{Revenue}^{\text{SPA}}(r) = \frac{1}{T}\sum_{t=1}^{T} 1(b_t^{(1)} > r)\max\{b_t^{(2)}, r\} \tag{6}$$

BIN-TAC is harder, as an agent's equilibrium decision to take the BIN option depends on their beliefs about the distribution of rival valuations. From the model, advertiser behavior is characterized by a threshold value $\overline{v}_j = \overline{v}_j(p, d, r)$ for each product, above which they will take the BIN option, and below which they will TAC. From Theorem 1, this threshold solves the implicit equation $\overline{v}_j - p = \frac{1}{d}E[\overline{v} - Y^*|Y^1 < \overline{v}_j]$, where $Y^* = \max\{Y^d, r\}$ and Y^1 and Y^d are the 1st and d-th order statistics of rival bids on product j. To solve this equation for fixed (p, d, r), we need to estimate the expected TAC payment $E[Y^*|Y^1 < s]$ for varying s.

Under symmetry, the joint distribution of valuations is exchangeable, and so the joint distribution of rival bids is exactly the same as the joint distribution of $N-1$ randomly selected bids. So our estimate of the TAC payment conditional on winning on product j is given by:

$$\text{TAC Payment}(s, r) = \frac{1}{T}\sum_{t=1}^{T} \frac{\sum_k 1(b_t^{(1)} < s)\max\{b_t^{(d)}, r\}}{\sum_k 1(b_t^{(1)} < s)}$$

where k indexes the N choices of $N-1$-length bid vectors for each auction, including zeros for bidders that didn't participate and restricting the sample only to product j.[21] We can then

the data generating process. However given our sample size, the bias this introduces is likely to be small.

[20] We use 100 bootstrap samples (i.e. samples of T impressions drawn randomly with replacement).

[21] It is correct to include the non-participating bidders, as in principle all N bidders are present in every period and so the distribution of rival bids drops only one of them — probably a bidder who would not have participated in any case.

solve for the equilibrium $\overline{v}(p,d,r)$ for each set of BIN-randomization parameters (p,d,r), and get a revenue estimate as follows:

$$\begin{aligned}\text{Revenue}^{\text{BIN-TAC}}(p,d,r) = &\frac{1}{T}\sum_{t=1}^{T} 1(b_t^{(2)} \geq \overline{v}(p,d,r))b_t^{(2)} + \frac{1}{T}\sum_{t=1}^{T} 1(b_t^{(1)} \geq \overline{v}(p,d,r) > b_t^{(2)})p \\ &+ \frac{1}{T}\sum_{t=1}^{T} 1(b_t^{(1)} < \overline{v}(p,d,r)) \sum_{j=1}^{d} 1(b_t^{(j)} \geq r) \max\{b^{(d+1)}, r\}\end{aligned} \tag{7}$$

Bundling Mechanisms: As we do not observe all the impression characteristics provided to advertisers in this market, we cannot consider the optimal bundling strategy. But we can consider bundling by product and user region, where the platform strips away all other user characteristics except for the region, so that advertisers are buying a random impression of a given size, on a given website, being viewed by a user from a particular US state. This is unlikely to be optimal, but provides a lower bound on the revenues from the bundling strategy.

For this analysis, we allow for bidder valuations to be asymmetric and vary by both product and region. Our estimate of a bidder's willingness to pay for this "generic impression" is just their average bid across all auctions of this product-region combination, taking their implicit bids when they didn't participate as equal to zero. Given that participation costs are zero and all bidders have strictly positive mean valuations, in the counterfactual world all bidders will participate in all auctions. We assume that these impressions are sold by second-price auction without reserve (since the bundling creates thick markets, a reserve isn't necessary).

Robustness to Informational Assumptions: The above theory and structural estimation follows the empirical auctions literature in treating bidder's valuations as private information.[22] A different modeling approach was suggested in an influential paper by Edelman, Ostrovsky and Schwartz (2007). They proposed a complete information model of sponsored search auctions. Their logic was that since these players compete with high frequency and can potentially learn each others' valuations, a complete information model may be a better approximation to reality than an incomplete information model.

[22]See for example Laffont and Vuong (1996). See also Athey and Nekipelov (2010) for a model of sponsored search models in this tradition.

Following this intuition, we also consider counterfactual simulations under complete information. The only model this affects is the BIN-TAC model, as under weak refinements the SPA equilibria under incomplete and complete information coincide. However in the BIN-TAC model we unfortunately now have multiple equilibria.[23]

To see this, consider a case where the bidder with the highest valuation is going to take the BIN option. Then the remaining bidders are indifferent between BIN and TAC, since in either case they will lose the auction and get payoff 0. We employ a trembling hand perfection refinement to eliminate this multiplicity. Specifically, for any probability $\epsilon > 0$ that the highest bidder will take the TAC option instead, the second-highest bidder faces a non-trivial choice between BIN and TAC. Applying this logic restores a generically unique equilibrium prediction.[24] We can therefore solve for the unique trembling hand perfect equilibrium of each auction, and estimate the expected revenues from the average sample revenues at any parameter vector.

We also perform a worst-case analysis over all rationalizable beliefs about rival strategies. From the point of view of revenue, the worst-case for BIN-TAC occurs when agents are least inclined to take the BIN option: specifically, when they believe that all other agents will choose to TAC and then bid zero. This implies that incentives to take the BIN option must be provided directly by the design, through the randomization parameter d and the reserve price r in the TAC auction. Since these beliefs are identical across all auctions, we can compute the indifference threshold $\overline{v}(p, d, r)$ implied by these beliefs, and then calculate revenue in exactly the same way as in the incomplete information case.

4.3 Results

The results are in Tables 4 and 5. We find that the optimal reserve when running a second price auction is high: nearly twice as high as the second highest bid. By contrast, BIN-TAC always uses relatively low reserves (all well below the average bid), and instead threatens to randomize among 3-4 bidders in order to get agents to take the high buy price (which is close in magnitude to the optimal SPA reserve). Interestingly, in the worst-case scenario the platform has to threaten randomization among 4 agents to get bidders to take BIN, since bidder beliefs are such that the TAC auction looks relatively attractive.

[23] This arises also in the generalized second price auction — see Edelman et al. (2007) and Varian (2007).
[24] We prove this in the supplementary appendix.

The welfare performance of these mechanisms is detailed in Table 5. The SPA without reserve earns revenue of 0.98 per auction, and leaves substantial consumer surplus — on average 1.97 per auction.[25] Adding the large optimal reserve improves revenue slightly (to 1.03 per auction), but hurts consumer surplus substantially (it falls to 1.44).

BIN-TAC does better than both of these mechanisms in terms of revenue. Interestingly, the consumer surplus is higher than under the SPA with targeting and reserve, implying BIN-TAC dominates the SPA in terms of both revenue and consumer surplus. This happens because the optimal SPA reserve price is very high — to extract revenue from the long right tail — and so many impressions are not sold, resulting in inefficiency and lower total welfare. By contrast, BIN-TAC has the BIN price to extract this revenue, and so the reserve is much lower, and more impressions are sold. Even accounting for distortions owing to the TAC auction, this is a welfare improvement.

By contrast, the bundling strategy underperforms. Revenues are much lower than the simple SPA, and consumer surplus falls even more dramatically. This is because there is considerable variation in match surplus across impressions even after conditioning on product and region, and so bundling along only these two dimensions destroys a lot of surplus.

Finally, the BIN-TAC results in the bottom part of the table show that the revenue estimates are relatively robust to how we model the information structure. However in models where the bidders are more informed, or dubious about the BIN option, consumer surplus is lower. In those cases the BIN decision is taken less often, thereby increasing the distortion from TAC auctions.

5 Conclusion and Future Work

We have introduced the BIN-TAC mechanism, designed to allow sellers to capture the surplus created by providing match information. This mechanism outperforms the second-price auction mechanism in this setting, and is preferable to bundling goods together by withholding information, at least when there is a reasonable size population of potential bidders. Moreover, we demonstrated through an example that the mechanism can closely approximate Myerson's optimal mechanism with ironing, despite its relative simplicity.

[25]The per auction revenue of 0.98 is lower than the average second highest bid of 1.07 in Table 2 because of a small fraction (2.3%) of auctions with only a single bidder, which will realize zero revenue in an SPA without reserve.

Our analysis of the exchange marketplace revealed that it has many features that make it a good place to apply our mechanism: large differences between the highest and second highest bid, and evidence of matching on user characteristics that the platform has chosen to make available to advertisers. Although the market does not fit our stylized model, we found that the BIN-TAC mechanism would nonetheless improve revenues and consumer surplus relative to the existing mechanism, a second price auction with reserve.

Due to data limitations we were not able to compare our mechanism to an optimal bundling strategy. Instead, we looked at what would happen if the platform only provided advertisers with product and user location information, rather than more detailed demographics. This bundling strategy performed poorly, but it is an interesting and open research question as to whether switching mechanisms to BIN-TAC is in fact better than retaining the SPA with a more clever bundling strategy.

References

Abraham, Ittai, Susan Athey, Moshe Babioff, and Michael Grubb, "Peaches, Lemons and Cookies: The Impact of Information Structure on Revenue in Common Value Auctions," 2010. working paper, Microsoft Research.

Adams, William J. and Janet Yellen, "Commodity Bundling and the Burden of Monopoly," *Quarterly Journal of Economics*, 1976, *90* (3), 475–498.

Athey, Susan and Denis Nekipelov, "A Structural Model of Sponsored Search Advertising Auctions," 2010. Working Paper, Harvard University.

Bergemann, Dirk, Alessandro Bonatti, and Maher Said, "Product Design in Ad Exchanges," 2011. Working paper, MIT Sloan.

___ **and** ___ , "Targeting in Advertising Markets: Implications for Offline vs. Online Media," 2010. Cowles Foundation Discussion Paper 1758R.

___ **and Maher Said**, "Dynamic Auctions: A Survey," *Wiley Encyclopedia of Operations Research and Management Science*, 2010.

___ **and Martin Pesendorfer**, "Information Structures in Optimal Auctions," *Journal of Economic Theory*, 2007, *137*, 580–609.

Chawla, Shuchi, Jason D. Hartline, David L. Malec, and Balasubramanian Sivan, "Multi-parameter mechanism design and sequential posted pricing," in "Proceedings of the 42nd ACM Symposium on Theory of Computing (STOC)" 2010, pp. 311–320.

Chu, Sean, Phillip Leslie, and Alan Sorensen, "Bundle-size Pricing as an Approximation to Mixed Bundling," *American Economic Review*, 2011, *101* (1).

Courty, Pascal and Hao Li, "Sequential Screening," *The Review of Economic Studies*, 2000, *67* (4), pp. 697–717.

Edelman, Benjamin, Michael Ostrovsky, and Michael Schwartz, "Internet Advertising and the Generalized Second Price Auction: Selling Billions of Dollars of Keywords," *American Economic Review*, 2007, *97* (1), 242—259.

Even-Dar, Eyal, Michael J. Kearns, and Jennifer Wortman, "Sponsored Search with Contexts," in "Internet and Network Economics, Third International Workshop (WINE)" 2007, pp. 312–317.

Ghosh, Arpita, Hamid Nazerzadeh, and Mukund Sundararajan, "Computing Optimal Bundles for Sponsored Search," in "Internet and Network Economics, Third International Workshop (WINE)" 2007, pp. 576–583.

Goldfarb, Avi and Catherine Tucker, "Online Display Advertising: Targeting and Intrusiveness," *Marketing Science*, 2011, *30* (3), 389—404.

___ **and** ___ , "Privacy Regulation and Online Advertising," *Management Science*, 2011, *57* (1), 57—71.

Laffont, Jean-Jacques and Quang Vuong, "Structural Analysis of Auction Data," *The American Economic Review*, may 1996, *86* (2), 414–420.

Levin, Johathan and Paul Milgrom, "Online Advertising: Heterogeneity and Conflation in Market Design," *American Economic Review, Papers and Proceedings*, 2010, *100* (2).

Lewis, Randall A and David H Reiley, "Does Retail Advertising Work? Measuring the Effects of Advertising on Sales via a Controlled Experiment on Yahoo!," 2011. Yahoo! Research Working Paper.

Lewis, Tracy and David Sappington, "Supplying Information to Facilitate Price Discrimination," *International Economic Review*, 1994, *35* (2), 309–327.

McAfee, Preston, "The Design of Advertising Exchanges," *Review of Industrial Organization*, 2011, *39* (3), 169—185.

___ , John McMillan, and Michael D. Whinston, "Multiproduct Monopoly, Commodity Bundling, and Correlation of Values," *Quarterly Journal of Economics*, 1989, *104* (2), 371—384.

McAfee, R. Preston, Kishore Papineni, and Sergei Vassilvitskii, "Maximally Representative Allocations for Guaranteed Delivery Advertising Campaigns," 2010. Yahoo! Research Working Paper.

Milgrom, Paul and Chris Shannon, "Monotone Comparative Statics," *Econometrica*, 1994, *62* (1), 157–180.

Muthukrishnan, S., "Ad Exchanges: Research Issues," 2010. manuscript.

Myerson, Roger B., "Optimal Auction Design," *Mathematics of Operations Research*, 1981, *6* (1), 58—73.

Ostrovsky, Michael and Michael Schwarz, "Reserve Prices in Internet Advertising Auctions: A Field Experiment," 2009. Working Paper.

Rochet, Jean-Charles and Philippe Chone, "Ironing, Sweeping, and Multidimensional Screening," *Econometrica*, July 1998, *66* (6), 788–826.

Varian, Hal, "Position Auctions," *International Journal of Industrial Organization*, 2007, *25* (7), 1163–1178.

Appendix

Proof of Theorem 1

Let a be a binary choice variable equal to 1 if the agent takes BIN and zero if TAC. Fix a player i, and fix arbitrary measurable BIN strategies $a_j(v)$ for the other players. Let q be the probability that no other agent takes the BIN option, equal to $\prod_{j \neq i} \left(\int 1(a_j(v) = 0) dF(v) \right)$.

Let $\pi(a, v)$ be the expected payoff to action a for type v given that the agent bids their valuation in any auction that follows. Then we have that $\frac{\partial}{\partial v}\pi(1, v) \geq q$, as a marginal increase in type increases the payoff by the probability of winning, which is lower bounded by q when taking the BIN option. Similarly we have that $\frac{\partial}{\partial v}\pi(0, v) \leq \frac{q}{d}$, as the probability of winning when taking-a-chance is bounded above by q/d. Then $\pi(a, v)$ satisfies the strict single crossing property in (a, v); it follows by Theorem 4 of Milgrom and Shannon (1994), the best response function must be strictly increasing in v, which in this case implies a threshold rule. It follows that any symmetric equilibrium must be in symmetric threshold strategies. So fix an equilibrium of the form in the theorem, and let the payoffs to taking taking BIN be $\pi_B(v)$ and to TAC be $\pi_T(v)$. They are given by:

$$\pi_B(v) = E\left[1(v > Y^1 > \overline{v})(v - Y^1)\right] + E\left[1(Y^1 < \overline{v})(v - p)\right]$$
$$\pi_T(v) = \mathbb{E}\left[1(Y^1 < \overline{v})1(Y^* < v)\frac{1}{d}(v - Y^*)\right]$$

The threshold type $\overline{v}$ must be indifferent, so

$$\begin{aligned} \pi_B(\overline{v}) &= \mathbb{E}\left[1(Y^1 < \overline{v})(\overline{v} - p)\right] \\ &= \mathbb{E}\left[1(Y^1 < \overline{v})\frac{1}{d}(\overline{v} - Y^*)\right] = \pi_T(\overline{v}). \end{aligned} \tag{8}$$

Next, we show a $\overline{v}$ satisfying Eq. (1) exists and is unique. The right hand side of Eq. (1) is a function of $\overline{v}$ with first derivative $\frac{1}{d}(1 - \frac{\partial}{\partial \overline{v}}\mathbb{E}[Y^*|Y^1 < \overline{v}]) < 1$. Since at $\overline{v} = 0$ it has value $p > 0$ and globally has slope less than 1, it must cross the 45° line exactly once. Thus there is exactly one solution to the implicit Eq. (1).

Proof of Theorem 2

By assumption, $\psi(v)$ single-crosses zero exactly once from below on $[\underline{\omega}_L, \overline{\omega}_L]$, so the implicit equation for r^* has exactly one solution. We next show that this first order condition is necessary. Fix d and $\overline{v} > p \geq r$ and define $p(r)$ implicitly as the BIN price that holds $\overline{v}$ constant as r changes. Then there are two effects of increasing the reserve r slightly: first, you can raise the BIN price without changing $\overline{v}$; second, if all bidders TAC, increasing the reserve raises the expected payment of some types, while decreasing the probability of sale.

The marginal increase in revenue from BIN auctions is:

$$nF(\overline{v})^{n-1}(1-F(\overline{v}))\frac{1}{d}\Pr(Y^d \leq r)$$

With probability $F(\overline{v})^n$ there are no BIN bidders. Writing $F_{\overline{v}}$ for $F(v|v<\overline{v})$, revenue from the TAC auction is given by:

$$F(\overline{v})^n \frac{1}{d}\sum_{k=1}^{d}\left[\sum_{j=k}^{d}\binom{n}{j}(1-F_{\overline{v}}(r))^j F_{\overline{v}}(r)^{n-j} r \right.$$
$$\left. + \int_r^{\overline{v}} \frac{n!}{d!(n-1-d)!} f_{\overline{v}}(s) F_{\overline{v}}(s)^{n-d-1}(1-F_{\overline{v}}(s))^d ds\right]$$

Taking a first order condition in r, canceling telescoping terms and simplifying:

$$F(\overline{v})^n \frac{1}{d}\sum_{k=1}^{d}\binom{n}{k} k(1-F_{\overline{v}}(r))^{k-1}F_{\overline{v}}(r)^{n-k}\left(1-F_{\overline{v}}(r)-rf_{\overline{v}}(r)\right)$$

Summing both marginal effects and expanding $P(Y^d \leq r)$:

$$n(1-F(\overline{v}))\left(\sum_{k=0}^{d-1}\binom{n-1}{k}(1-F_{\overline{v}}(r))^k F_{\overline{v}}(r)^{n-1-k}\right) +$$
$$F(\overline{v})\sum_{k=1}^{d}\binom{n}{k} k(1-F_{\overline{v}}(r))^{k-1}F_{\overline{v}}(r)^{n-k}\left(1-F_{\overline{v}}(r)-rf_{\overline{v}}(r)\right)$$

Changing summation limits, factorizing, eliminating constants and setting the FOC = 0:

$$(1-F(\overline{v}))+\left(1-F_{\overline{v}}(r)-rf_{\overline{v}}(r)\right)F(\overline{v})=0$$

Now since $F_{\overline{v}} = F(v|v<\overline{v}) = F(v)/F(\overline{v})$, we can simplify and solve to get $r^* = \frac{1-F(r^*)}{f(r^*)}$.

Next, the optimal BIN price $p > r^*$ must be such that $\overline{v} \geq \underline{\omega}_H$ (only high types take the BIN option). Let $p(\overline{v}, r)$ and $R(\overline{v}, r)$ be defined as in the text. There are three effects of a marginal increase in $\overline{v}$. First, the second highest bidder may have valuation $\overline{v}$ and choose not to take BIN, which decreases revenue by $\overline{v} - p(\overline{v}, r)$. The probability of $V^2 = \overline{v}$ is given by $n(n-1)f(\overline{v})(1-F(\overline{v}))F(\overline{v})^{n-2}$. The second is that that highest bidder may have valuation $\overline{v}$ and choose not to take BIN, reducing revenue by $p(\overline{v}, r) - R(\overline{v}, r)$. This happens

with probability $nf(\overline{v})F(\overline{v})^{n-1}$. Finally, the highest bidder may have valuation above $\overline{v}$ and the second highest below it, which raises revenue by $\frac{\partial p(\overline{v},r)}{\partial \overline{v}}$. This happens with probability $n(1-F(\overline{v}))F(\overline{v})^{n-1}$. Setting the sum of these effects equal to zero, evaluating the expression at r^* and eliminating common factors we get:

$$f(\overline{v})\left(((n-1)(1-F(\overline{v}))(\overline{v}-p(\overline{v},r^*))+F(\overline{v})(p(\overline{v},r^*)-R(\overline{v},r^*))\right)=(1-F(\overline{v}))F(\overline{v})\frac{\partial p(\overline{v},r^*)}{\partial \overline{v}}$$

Proof of Theorem 3

We prove each of the results in turn. For part (i), we construct a BIN-TAC mechanism that achieves exactly the same outcomes as SPA-T for any type realization. Let the SPA-T have optimal reserve r^*, and let the BIN-TAC mechanism have TAC reserve r^*, randomization parameter $d=1$ and BIN price $p=\overline{\omega}_H$. Then no type will take the BIN option in equilibrium (it is strictly dominated), and so the TAC auction will always occur. Since $d=1$, this is just an SPA with reserve r^*. Since for this particular choice of parameters BIN-TAC does as well as SPA-T, in general BIN-TAC dominates SPA-T. Strict dominance follows from the uniform example in the text.

For part (ii), we will argue that there is an open interval $(\underline{\beta},1]$ on which SPA-T dominates SPA-B, and therefore by part (i), so does BIN-TAC. First we argue that when $\beta=1$, SPA-T does better than SPA-B. Since the objective is surplus maximization, the optimal SPA-T and SPA-B share a common reserve of zero. Total surplus is just the valuation of the winning bidder. Under SPA-T this is obviously maximized, since in the SPA-T bidders bid their valuation, and the highest bidder wins. Under SPA-B this needn't be maximized, since bidders bid their expected valuation, and the highest bidder wins. So whenever the bidder with the highest realized valuation did not have the highest expected valuation, the highest valuation bidder does not win. Since these events happen with positive probability, SPA-T achieves higher expected payoff than SPA-B for $\beta=1$. Now both $\pi(\text{SPA-T},r,\beta)$ and $\pi(\text{SPA-B},r,\beta)$ are continuous in r and β. This holds by inspection for β, and is true for r because the type distributions are atomless. Then a triangle inequality argument suffices to extend the result at $\beta=1$ by continuity to an interval $(\underline{\beta},1]$ for the functions $\max_r \pi(\text{SPA-T},r,\beta)$ and $\max_r \pi(\text{SPA-B},r,\beta)$.

For part (iii), consider a BIN-TAC mechanism with $d=n$, and a reserve of v_L. The highest BIN price that makes electing BIN optimal for high types is $p=v_H-\frac{(v_H-v_L)}{n}=\frac{n-1}{n}v_H+\frac{1}{n}v_L$.

Then for any $\beta \in [0,1]$, we have:

$$\begin{aligned}\max_{\theta} \pi(\text{BIN-TAC}, \theta, \beta) &\geq \left(1-(1-\alpha)^n - n\alpha(1-\alpha)^{n-1}\right) v_H + n\alpha(1-\alpha)^{n-1}\left(\beta v_H + (1-\beta)p^{BIN}\right) \\ &\quad + (1-\alpha)^n v_L \\ &= \left(1-(1-\alpha)^n - n\alpha(1-\alpha)^{n-1}\right) v_H + n\alpha(1-\alpha)^{n-1}\left((1-\frac{1-\beta}{n})v_H + \frac{1-\beta}{n}v_L\right) \\ &\quad + (1-\alpha)^n v_L \\ &\geq \left(1-(1-\alpha)^n - \alpha(1-\alpha)^{n-1}\right) v_H + \left(\alpha(1-\alpha)^{n-1} + (1-\alpha)^n\right) v_L\end{aligned}$$

where the last step uses the fact that the minimum of the function occurs at $\beta = 0$. Under the SPA-B, all types bid the same, there is no consumer surplus, and revenue is equal to:

$$\max_{\theta} \pi(\text{SPA-B}, \theta, \beta) = \alpha v_H + (1-\alpha) v_L$$

The final line of both expressions is a weighted average of v_L and v_H; it suffices to show that the mass on v_L is lower under BIN-TAC. This requires $(\alpha(1-\alpha)^{n-1} + (1-\alpha)^n) < (1-\alpha)$. After a bit of simple algebra, this is equivalent to showing $(1-\alpha)\left(1-\alpha(1-\alpha)^{n-2} - (1-\alpha)^{n-1}\right) \geq 0$, which holds by binomial expansion of 1 with equality for $n = 2$ and strictly for $n > 2$. This proves dominance; strict dominance follows by noting that the BIN-TAC payoff is strictly higher for $\beta > 0$.

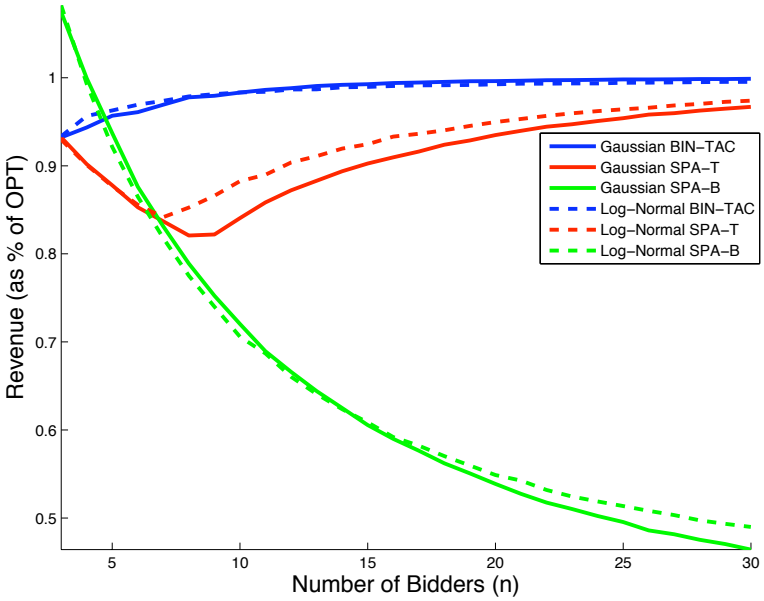


Figure 5: **Revenue Performance vs Number of bidders.** Simulated expected revenues for different mechanisms as the number of bidders n varies, in an environment where F_L has mean 1 and standard deviation 0.5, the match probability is 0.05 and the match increment is 5.

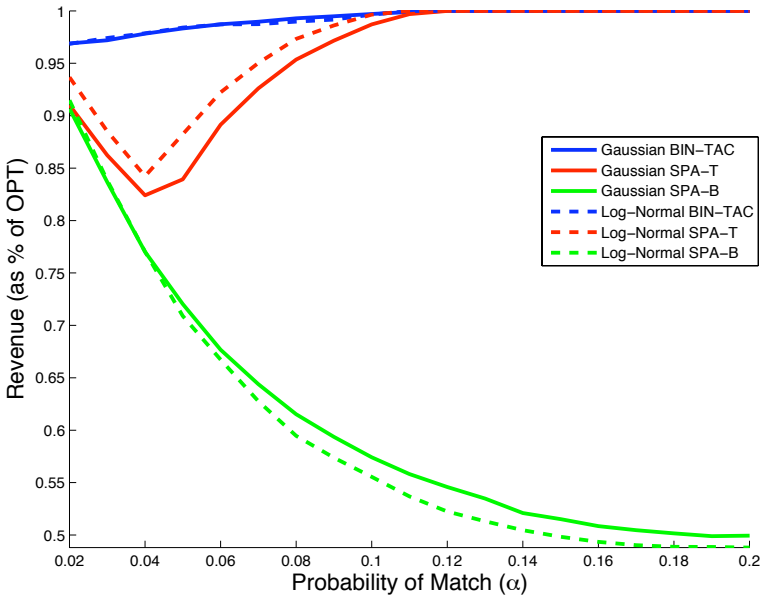


Figure 6: **Revenue Performance vs Match Probability.** Simulated expected revenues for different mechanisms as the probability of a match α varies, where F_L has mean 1 and standard deviation 0.5, the number of bidders is 10 and the match increment is 5.

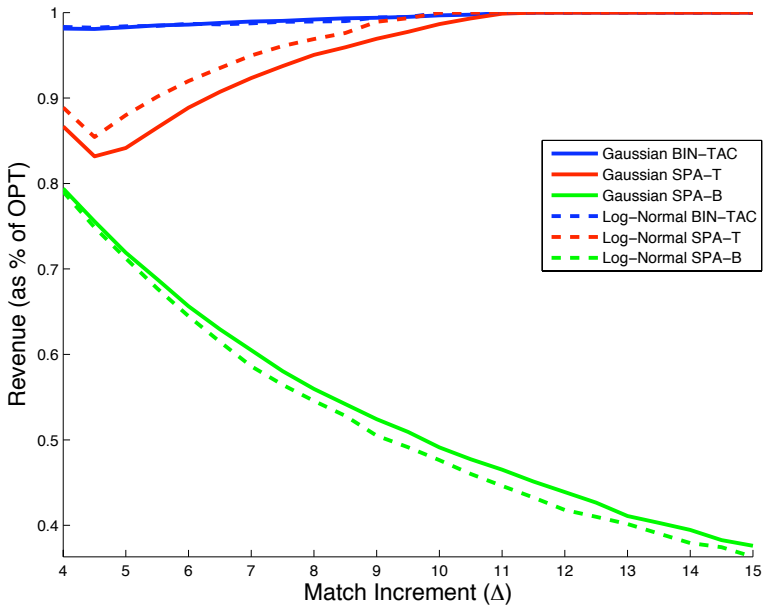


Figure 7: **Revenue Performance vs Match Increment.** Simulated expected revenues for different mechanisms as the match increment Δ varies, where F_L has mean 1 and standard deviation 0.5, the match probability is 0.05 and the number of bidders is 10.

Table 2: Summary Statistics: Microsoft Advertising Exchange Display Ad Auctions

	Mean	Median	Std. Dev.	Min	Max
Bid-Level Data					
Average bid	1.000	0.565	2.507	0.0000157	130.7
Number of bids	508036				
Auction-Level Data					
Winning bid	2.957	1.614	5.543	0.00144	130.7
Second highest bid	1.066	0.784	1.285	0.00132	39.22
Number of bidders	6.083	6	2.970	1	15
Bid correlation	0.01				
Number of auctions	83515				
Advertiser-Level Data					
% of auctions participated in (p1)	0.697	0.0251	4.641	0.00120	88.28
% of auctions won if participated (p2)	38.90	29.59	35.50	0	100
Correlation of (p1,p2)	-0.09				

Summary statistics for the full dataset, which is a 0.1 percent sample of a week's worth of auction data sampled within the last two years. An observation is a bid in the top panel; an auction in the middle panel; and an advertiser in the last panel. Bids have been normalized so that their average is 1, for confidentiality reasons. The bid correlation is measured by selecting a pair of bids at random in every auction with at least two bidders, and computing the correlation coefficient.

Table 3: Matching on Region

	Participation		Bids	
Advertiser Website Pageview Ratio	0.029	0.329***	0.264***	0.286***
	(0.022)	(0.015)	(0.052)	(0.053)
Time-of-Day Fixed Effects	yes	yes	yes	yes
Product-Region Fixed Effects	yes	yes	yes	yes
Advertiser Fixed Effects	no	yes	no	yes
N	5581749	5581749	417557	417557
R^2	0.02	0.34	0.04	0.26

Results from OLS Regressions. In the first two columns, the dependent variable is a dummy for participation. The sample used in the regressions consists of all auction-bidder pairs, limited to the 10% of bidders who participate most often. In the last two columns, the dependent variable is the bid. The sample used in the regressions only includes bids from the 10% of bidders who bid most often. The independent variable is the population-weighted fraction of pageviews of the advertiser's website that come from the region the user is in. Time-of-day fixed effects refer to a dummy for each quarter of the day, starting at midnight. Product-region fixed effects are dummies for the page-group advertised on, and the state the user is located in. Standard errors are robust. Significance levels are denoted by asterisks (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

Table 4: Optimal Parameter Choices

Policy	p	d	r
SPA-T	-	-	1.96
BIN-TAC (incomplete information)	2.60	3	0.43
BIN-TAC (complete information)	1.95	3	0.65
BIN-TAC (rationalizable worst case)	2.10	4	0.65

Revenue-maximizing parameter choices. For each of the above mechanisms, we find these by maximizing the revenue functions defined in the main text over the available parameters numerically using a grid search.

Table 5: Counterfactual Revenues and Welfare

Policy	Revenue	Consumer Surplus	Total Welfare
SPA-T (no reserve)	0.983	1.974	2.957
	(0.004)	(0.019)	(0.019)
SPA-T (optimal reserve)	1.028	1.471	2.499
	(0.005)	(0.018)	(0.020)
BIN-TAC	1.075	1.633	2.708
	(0.005)	(0.018)	(0.020)
SPA-B	0.644	0.730	1.374
(bundling by product-region)	(0.006)	(0.016)	(0.015)
Robustness to Informational Assumptions			
BIN-TAC (complete information)	1.072	1.589	2.661
	(0.005)	(0.018)	(0.020)
BIN-TAC (rationalizable worst case)	1.066	1.530	2.596
	(0.005)	(0.018)	(0.020)

Counterfactual simulations of average advertiser revenues, consumer surplus and total welfare (sum of producer and consumer surplus). All statistics reported outside parentheses are averages across impressions; those in parentheses are standard errors computed by bootstrapping the full dataset (i.e. they reflect uncertainty over the true DGP). Six different simulations are run. The first is of a second price auction without reserve, while the second is of a second price auction with optimal (revenue-maximizing) reserve. The third is of the BIN-TAC mechanism, under the incomplete information structure outlined in the text. The fourth is a bundling counterfactual where the impressions are bundled according to the product (i.e. URL and ad size) and user region, and sold by second-price auction. The last two are robustness checks, varying the informational assumptions made for BIN-TAC. In the complete information case, bidders know the valuations of the other participants, and made BIN decisions accordingly. In the rationalizable worst-case model, bidders assume they will only have to pay the reserve price in TAC auction, and therefore take the BIN option more rarely. Where applicable, the parameters used are the optimal parameters from Table 4.

ISSN 2042-2695

CENTRE *for* ECONOMIC PERFORMANCE

CEP Discussion Paper No 1133

March 2012

Incentives for Quality over Time – The Case of Facebook Applications

Jörg Claussen, Tobias Kretschmer and Philip Mayrhofer

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Abstract
We study the market for applications on Facebook, the dominant platform for social networking and make use of a rule change by Facebook by which high-quality applications were rewarded with further opportunities to engage users. We find that the change led to quality being a more important driver of usage while sheer network size became less important. Further, we find that update frequency helps applications maintain higher usage, while generally usage of Facebook applications declines less rapidly with age.

Keywords: usage intensity, social media, platform management, two-sided markets
JEL Classification: L1, L50, O33

This paper was produced as part of the Centre's Productivity and Innovation Programme. The Centre for Economic Performance is financed by the Economic and Social Research Council.

Acknowledgements
We thank participants at the 2010 ZEW conference on platform markets, 2011 ZEW conference on online markets, the 4th INTERTIC conference on competition in high-tech markets, the 2011 annual conference of the Royal Economic Society, the 2nd workshop on the economics of ICTs in Evora, the 9th annual International Industrial Organization conference, the BRIE-ETLA conference on the changing competitive landscape in the ICT domain and its socio-economic implications, seminar participants at LMU Munich, Oliver Alexy, Annika Bock, David Genesove, Jose Luis Moraga Gonzalez, Stefan Wagner, and Tim Watts for helpful comments and discussions.

Jörg Claussen is a Postdoc at the Ifo Institute for Economic Research at the University of Munich. Tobias Kretschmer is a Professor of Management at the University of Munich, Department Head at the Ifo Institute and an Associate in the CEP's Productivity and Innovation Programme. Philip Mayrhofer is an entrepreneur and consultant.

Published by
Centre for Economic Performance
London School of Economics and Political Science
Houghton Street
London WC2A 2AE

1. Introduction

A platform sponsor faces a chicken-and-egg problem since platform markets typically display indirect network effects (Farrell and Klemperer 2007). Consumers will only use the platform if there are sufficient complementary goods available, while producers of complementary goods will only provide them if the number of potential users is sufficiently large (Caillaud and Jullien 2003; Evans 2003). When managing a platform (or multi-sided market), an important question is how to ensure sufficient supply of each side from independent suppliers. To achieve this, platform providers often open up their platform to third-party developers who supply additional modules and functionality (Wheelwright and Clark 1992; Tiwana et al. 2010). Opening one side of the market poses challenges to the platform owner as its ability to generate revenues and profits depends on the quality and quantity of both market sides, none of which the owner controls directly (Boudreau 2008; Boudreau and Hagiu 2009; Hagiu 2011).

Social media are transforming businesses in multiple ways. Specifically, users become increasingly interconnected and that communication influences their purchasing decisions (Godes et al. 2005; Tucker and Zhang 2011). In the context of multi-sided markets, this implies that a platform is not just an intermediary between singular, isolated consumers and a number of complementary products. Rather, due to social media tools and services, the adoption of complementary products by consumers is influenced by direct network effects (Katz and Shapiro 1985), bandwagon effects (Banerjee 1992; Dellarocas et al. 2010), or word-of-mouth (Katz and Lazarsfeld 1955; Dellarocas 2003; Godes and Mayzlin 2009). The ability to tap into social media channels affects developers' decisions of investing in a platform or not, making the management of platforms and multi-sided markets even more complex.

The managerial challenge is profound and matters for an increasing number of companies. On the one hand, social media services such as LinkedIn or Facebook initially offered a platform for users to interact and later opened this platform to third-party developers who build and market complementary services.[1] On the other hand, two-sided markets such as auction, shopping and other e-commerce platforms are also beginning to heavily integrate social media to enable user interaction.[2] Second, managing social media integration in multi-sided markets is an ongoing process rather than a one-off optimization problem. Social media facilitates fads and bandwagon processes and actions by the platform operator and third-party developers must be constantly monitored to prevent dynamics damaging one market side in a way that makes it unattractive for the other market side to participate.

We study Facebook, the largest social network on the Internet and a platform operator for add-on programs, in the initiation phase, which is crucial in the lifecycle of a multi-sided market. The market for Facebook applications was highly dynamic in the early stages, but suffered from a flood of low-quality applications, which was detrimental to user experience. Facebook undertook a number of changes to address this. Specifically therefore, we ask if a change in the rules by which Facebook applications can attract and engage users in February 2008 changed developers' incentives for quality provision for applications and ultimately the drivers of what makes for a successful application.

[1] Social networking platforms act as two-sided markets as soon as they rely on advertising to finance their operations. However, in this paper this side is less important as advertisers generally do not engage directly with the other side by offering products consumed on the platform itself (Evans 2008).

[2] Services using social media for shopper coordination are often termed social commerce services.

Managing a platform has the ultimate goal of maximizing monetization opportunities. As revenues for platform owners are often generated through advertising or transaction-based charges, managing usage intensity or frequency (which in turn increases advertising effectiveness) is often at the core of platform management. This may be done through non-price instruments imposing rules and constraints, creating incentives and shaping demand and supply behavior (Boudreau and Hagiu 2009). Hagiu (2011) identifies a quality/quantity tradeoff, since both higher quality and quantity are attractive to consumers, but higher quality presents an entry barrier to complementary goods providers, thus reducing quantity. Casadesus-Masanell and Halaburda (2011) argue that platform owners may limit the number of applications on a platform to realize larger benefits from application-level direct network effects. We contribute to the literature by showing that increased platform quality can be achieved with "soft" quality incentives and no hard exclusion of low-quality participants is necessary.[3] We also argue and demonstrate empirically that a change in "soft" incentives changes other drivers of application success.

We study applications developed for Facebook and observe their usage between September 2007 and June 2008.[4] On Facebook, the amount of information an application can send out to users critically influences usage intensity. In February 2008, Facebook implemented a rule change regarding the amount of notifications applications could send out: before February 2008, all applications could send out the same amount of messages, while thereafter the amount of notifications permitted was determined by how frequently the notifications were clicked on, a useful proxy for an application's ability to attract and retain users. This increased incentives for producing high-quality applications and punished applications that send out information deemed useless or even annoying by users. To isolate the effects of this change, we focus on a twenty-week time window around the change. We use this change (assumed to be endogenous to the platform operator but exogenous for application developers) to analyze how potential drivers of usage intensity changed in response. This natural experiment-like change in the effectiveness of word-of-mouth channels therefore allows for a similar identification as for field experiments (Aral and Walker 2011; Goldfarb and Tucker 2011; Animesh et al. 2011).

We use a rich, longitudinal data set on 7,784 applications on the social networking site Facebook. This setting is useful for several reasons. First, we have data on applications soon after the launch of the platform, which lets us examine the dynamics of a nascent and dynamic market. Second, our complete listing of applications on the platform avoids selection and survivor biases. Third, Facebook is one of the largest and most successful platforms for applications, making it relevant for the entire industry.

We estimate random- and fixed-effect OLS models and analyze how the drivers of usage intensity are affected by the rule change. We find that application quality matters more for usage intensity after the change, in line with Facebook's stated goals of the change. Conversely, we find that the rule change led to quantity (as expressed by the number of installations) becoming less important. The frequency with which applications are updated (a proxy for the degree to which an application is managed and maintained) gains in importance as a driver for usage intensity. Further, while usage intensity always declines as applications become older the decline is less severe after the rule change, which implies that the intervention was successful in keeping adopters engaged over time. Finally, the portfolio

[3] Evaluating quality of platform participants and excluding low-quality applications is costly and prone to error. Circumstantial evidence for this is given by regular debates revolving around applications rejected on Apple's market for iOS applications. Conversely, not imposing any quality restrictions may lead to a flooding of low-quality applications as observed in the 1983 Atari shock (Coughlan 2004).

[4] The application platform was opened in May 2007.

effect of belonging to a family of applications by the same developer becomes more positive after the change.

The paper proceeds as follows. We first present the industry context and describe Facebook's rule change. Thereafter, we discuss the economics of usage intensity as an indicator of application success. Our empirical model and results follow. We conclude with a discussion and an outlook for future work.

2. Industry Context

2.1. Applications on Facebook

Facebook is the major player in social networking websites (other examples are Google+ or LinkedIn). Consumers use social networking services to interact with friends, family members, and increasingly business partners. Core components include personal mini-homepages with which a user creates a digital representation of him-/herself (Boyd and Ellison 2007) as well as different means to communicate (personal messages, boards, chats) and to exchange different media.[5] Facebook is the largest and fastest-growing social network with over 800 million active users, of which 75% are outside the United States (as of November 2011).[6] In May 2007, Facebook launched an application platform which allows third parties to develop software that deeply integrates into the social network and enables additional service not currently covered by the core components of the social network. In May 2008, one year after the platform launched, more than 30,000 applications had been developed, with more than 900 million installations in total.

2.2. Entry of application developers

As in most markets with indirect network effects, platform operators want to encourage a wide variety of applications and experimentation in parallel (Church and Gandal 2000; Boudreau et al. 2008). Hence, they provide developers with a set of tools that decrease their development costs and thus entry barriers. This leads to high entry rates both from new entrants as well as from developers with multiple applications and affects both the users' experience and the developers' incentives. On the one hand, a large variety of applications presents novel challenges for consumers to discover and adopt applications (Oestreicher-Singer and Sundararajan 2006; Hervas-Drane 2010). On the other hand, high rates of entry could result in particularly high levels of competition, which in turn would diminish profits and incentives around the platform (Boudreau 2008).

Facebook encouraged entry of as many developers as possible. The company offered strategic subsidies to third-party developers (Shapiro and Varian 1998) by providing open and well-documented application programming interfaces, multiple development languages, free test facilities, as well as support for developers through developer forums and conferences. Facebook also has minimal requirements for applications to be included in the official directory and it does not "police" or discourage developers imitating or producing "copy-cats" of existing applications.

2.3 Adoption and usage of applications

Having a large variety of applications has important consequences for consumers' product search and adoption. On Facebook, adoption and usage takes place in a strongly embedded

[5] For example, Facebook is the largest online photo sharing utility.
[6] Source: http://www.facebook.com/press/info.php?statistics, accessed November 3rd 2011.

social context. The functionality provided by the platform operator lets developers build applications designed to intensify social interactions (Boudreau and Hagiu 2009).

Thus, application discovery and adoption is highly influenced by a user's social context. Users are passively influenced through the visibility of usage patterns such as reviews, ratings or matching mechanisms (Oestreicher-Singer and Sundararajan 2006; Hervas-Drane 2010). Active forms of social influence take the form of recommendations which are directly conveyed via predominantly digital or online word-of-mouth processes (Katz and Lazarsfeld 1955). IS and Marketing scholars have examined the conditions under which consumers are likely to rely on others' opinions in their purchase decisions, the motives for people to spread the word about a product, and the variation in strength of influence people have on their peers in word-of-mouth communications (Dellarocas 2003; Phelps et al. 2005; Bampo et al. 2008; Agarwal et al. 2009). It is widely acknowledged that in such contexts bandwagon processes – positive feedback loops where adoption decisions by some increase the incentive or pressure to adopt for others – are common (Katz and Shapiro 1985; Katz and Shapiro 1986; Abrahamson and Rosenkopf 1993).[7]

2.4 Monetization of applications

When Facebook launched its platform for third-parties in May 2007, developers may have been primarily intrigued by the opportunities to integrate their applications in Facebook's service. However, there was also an economic opportunity from revenues generated within the application. Importantly, Facebook decided not to take a share of transaction sales initially, leaving developers to capitalize on this revenue stream.[8] Facebook left it open to developers to monetize their application pages through advertising or other transactions that they control themselves. Also, Facebook deliberately did not impose restrictions on the form of advertising. The most common forms are advertisements next to the website's content and core functionality.

Facebook's objectives are largely aligned with the objectives of their third-party developers and rely on capitalizing their active user base. Revenues are realized via selling advertising space to brands, advertisers or Facebook applications that target specific users. Next to each application's canvas page (the space allocated to an application), Facebook can place its own advertising. As a consequence, the more users engage with applications, the more page impressions or time Facebook is able to sell to advertisers. Consequently, the level of revenue that can be realized is directly determined by the number of active users of the platform and applications.[9] Thus, growing the platform (applications) and keeping existing users active (and therefore generating transactions or looking at and clicking on ads) is among their most important objectives.

[7] Another feature relates to the costs that users incur in installing and using applications. Due to the dominant business model of indirect monetization, the vast majority of applications are free to use. Also, due to technical and design features, users can install and use multiple applications in parallel, thus "multi-home" (Rochet and Tirole 2003).

[8] Due to the (open) installation process and the lack of a payment system, Facebook could not take a revenue cut from developers without further development. In contrast, Apple takes a 30% revenue share from all sales in its iTunes store.

[9] Gnyawali et al. (2010) have shown a performance-increasing effect of opening up a platform to third-party codevelopment.

3. Facebook's Rule Change

Facebook users adopt applications through two main channels. First, users of an application can directly invite friends who are not currently users of the application (invites). Second, Facebook users get regular updates on friends' activities from the built-in "News Feed". To some extent, applications can send messages to this news feed and signal a friend's activity in this particular application (notifications).

Both channels have been influenced by Facebook. In the very first phase of the application platform (from launch in May to August 2007) invites and notifications could be sent almost without restrictions. Application developers used this to "spam" many of their users' friends. In September 2007 Facebook imposed a set of restrictions (the number of invites and notifications by user was limited). In the following months these rules remained unchanged.[10]

However, after months of steady growth, on February 6th 2008 Facebook announced a rule change such that notifications and invites would be allocated based on user feedback. Applications whose users react more heavily to notifications/invites that are sent out (a measure for relevance of the notifications/invites), would be able to send out more notifications/invites. One week later, feedback allocation was launched for notifications, requests, and invites. Facebook motivated this change by the expectation that the new system "provide[s] users with more compelling notifications and fewer notifications that they are likely to ignore or consider spam" (Figure 1). Further, they "hope this change incentivizes developers to improve the quality of their notifications".

Figure 1: Announcement of rule change on Facebook's developer blog[11]

Feedback-based allocations for notifications

By Tom Whitnah - Wednesday, February 6, 2008 at 9:47am

To improve the Facebook Platform user experience and to reward compelling applications, we will be rolling out a feedback-based system that allots notifications in proportion to user response. Applications will no longer have a static upper limit of 40 notifications per user per day. Instead the number of notifications per application will be based on a range of factors including the rates that users ignore, hide, and report notifications as spam.

The new system aims to provide users with more compelling notifications and fewer notifications that they are likely to ignore or consider spam. We hope this change incentivizes developers to improve the quality of their notifications and encourage their users to send notifications to interested friends.

Before this change goes out, we will be providing two new Insights statistics tabs. These tabs provide developers with their current per user notification threshold as well as metrics on notifications that they can use to understand how users are responding. While there is a general correlation between good response rates and higher thresholds, other factors and metrics will be used to determine these scores as well and the allocations will adjust themselves accordingly.

The new Insights tabs will be available later this week and users will start seeing changes next week. Please send your feedback to developers-help@facebook.com with [notifications allocations] in the subject field.

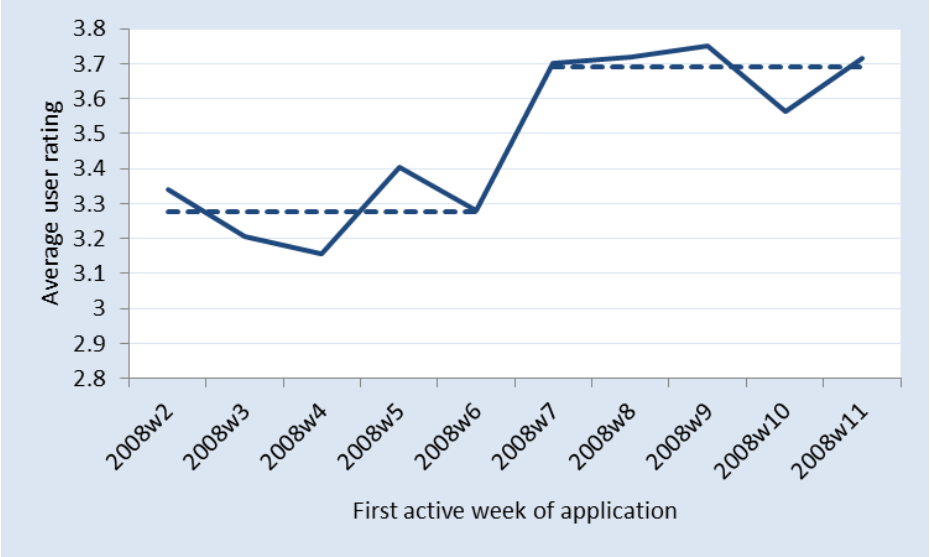
We want to understand how this rule change affected Facebook's market for applications. Did the rule change lead to the expected increased quality levels of applications? And was increased quality from then on a more effective driver of application usage? Finally, how were other drivers of application usage influenced by the rule change?

While the second and the third question are addressed in the following sections, we can already get insights on the first question, i.e. whether the rule change led to increased quality levels of applications. Figure 2 plots average quality ratings of Facebook applications

[10] To the best of our knowledge based on the official announcements of Facebook to its developers.
[11] Available at http://developers.facebook.com/blog/post/77, accessed September 13, 2011.

against their launch date.[12] We see that applications launched after the intervention (in the seventh week of 2008) immediately achieved a significantly higher average quality of around 0.4 points[13].[14] This is strong suggestive evidence that the rule change resulted in higher incentives for quality.

Figure 2: Average quality of applications launched before and after the rule change in 2008w7 (dotted line is average rating before/after the rule change)



So what could have motivated Facebook to initiate these changes? If application quality had always been Facebook's main goal it would be surprising to see no "hard" quality vetting process at the start of the applications market and if that was not feasible, not even a rule rewarding (notification and/or application) quality directly, but rather size through the number of notifications per user per day. While we cannot work out Facebook's aims (apart from the ones stated in the announcement) with certainty, the strategic subsidies to developers and the ease of getting applications listed suggests that inviting entry may have been on top of Facebook's agenda in the early stages of the applications market. This was reinforced through highly publicized success stories which created a gold rush among developers, e.g. the music application iLike grew to several million users within days. Within weeks, several thousand application developers had signed up for access credentials to the platform and had started to launch a wide variety of applications. For users, the myriad applications launched early on helped "educate" them about installing add-on applications to make the service fresh and exciting. Users learned quickly. Through invites and a flood of notifications in their news feed, the vast majority of users had installed at least one application within weeks. Also, many users installed dozens of applications at the same time (multi-homing is comparatively costless here), sometimes even several with largely identical functionality (e.g. within the first month there were several enhanced "walls" that allowed posting and exchanging multi-media items).

After the initial enthusiasm, user sentiment towards applications changed. With a larger installed base of applications and the increasing professionalization of developers in

[12] Data and variables are introduced in section 5.
[13] Measured on a five-point Likert scale.
[14] In this stage of the application market, applications have still been fairly simple and easy to program. This might explain why application quality reacted immediately to the enactment of the rule change, even though it has only been announced a couple of days earlier.

terms of exploiting the opportunities to use the "viral channels", the volume of notifications and invites grew exponentially. Users became annoyed by constant updates about their friend's activities and applications. For both Facebook as the platform operator and the developers this threatened to lead to adverse effects as instead of adopting and using applications, users would start ignoring notifications and requests.

Facebook's rule change came precisely during the time when notifications became annoying in the eyes of users. While the change did not increase entry barriers as such, it became more difficult for low-quality applications to gather large numbers of engaged users. Quality, as announced by Facebook itself, was rewarded more directly. Our econometric analysis looks at the actual effects to assess how this rule change affected the dynamics of applications usage.

4. Economics of Usage

4.1 Application quality and usage intensity

The stated aim of Facebook's rule change was to reward high-quality applications. Given the monetization opportunities discussed above, an application will likely be considered high-quality if it keeps users' interest and engagement. Therefore, the quality as perceived by users and ultimately reflected in monetization opportunities for application developers and Facebook itself will be closely linked to the ability to engage users. Hence, it is interesting to see how Facebook's rule change affected the relevant measure of application success. Finding the most relevant measure for a free-to-install application is not trivial. Conventional measures like profits or revenues are either not meaningful (as far as direct sales revenues are concerned) or not obtainable (as far as advertising revenues are concerned). The success of free software or goods more generally is therefore often measured in the number of users or installations. This may be misleading for at least two reasons: First, the number of installations (typically measured by the number of installation packs downloaded from the application's website) may double-count adopters if they repeatedly install software on different hardware or if they install upgrades to the basic version they already have. That is, the number of installations may be overstating the actual number of active copies of a piece of software – especially if it is frequently updated, requiring repeated installations. Second, even if the number of technically active copies is accurately measured, not all potential users will actually use it regularly. This is particularly relevant if a good comprises access and use elements, and most revenues arise from use, not access (Grajek and Kretschmer 2009). In the case of Facebook applications this would seem especially true as access is free (installing an application is costless) and usage increases the likelihood that a user will notice the advertisements from which the application derives its revenues. Hence, usage intensity is a useful metric for application success, which ultimately will translate to advertising revenues (unmeasured by us).

4.2 Drivers of usage intensity

We are interested in how Facebook's rule change may have affected the drivers of usage intensity in the market for applications. That is, we are interested in how the drivers of application success have changed after the rule change.

4.2.1 Application quality

Facebook's press release explicitly states the intention that "this change incentivizes developers to improve the quality of their notifications" and "to reward compelling notifications" (Figure 1). That is, notifications and users' application experience are linked

both through the new invitation process and the retention effect of in-app notifications. High-quality applications are more likely to be installed if a user receives an invitation, and they will be used more intensively if the notifications generated are meaningful. The notification process prior to the change introduced noise in this process by letting all applications issue the same number of notifications per user. Post-change, more successful applications could send out more notifications, leading to higher user engagement and average usage intensity.

Hence, we expect that the effect of quality on usage intensity will increase after Facebook's rule change.

4.2.2 Update activity

The frequency of updates is a choice variable by application developers. As Facebook applications usually have fairly limited functionality, an application can over time become less attractive for an individual who has already installed the application. However, user interest can be retained if an application is regularly updated and improved. Updating an application could include adding new features, new content, or just changing the application's appearance. Applications that are actively managed and updated regularly are therefore expected to better retain their customers and achieve higher usage intensity. Facebook's rule change was intended to improve user experience by making notifications more relevant, and notifications are more likely to be relevant if an application is managed proactively, thus leading to higher usage intensity.

Hence, the effect of upgrades on usage intensity is likely to increase after Facebook's rule change.

4.2.3 Installed base effects

The installed base of users of an application can relate to usage intensity through network effects as well as the composition of users at different points in time (Grajek and Kretschmer 2009; Cabral 2006). If there are network effects at the application level, more users should result in higher usage intensity. However, for local network effects, i.e. if an application becomes more attractive the more friends of a user have installed it, we do not expect a positive effect of an application's (global) installed base on usage intensity. The second mechanism affecting usage intensity through installed base is the composition of users over time. If adopters are heterogeneous and high-intensity users adopt early, *average* usage intensity is expected to decline with a growing installed base (Cabral 2006). Conversely, if diffusion is an epidemic process among users with similar preferences, we do not expect a positive effect of an application's installed base on usage intensity. Absent micro-(user-application-level) data, we cannot disentangle these effects empirically, but we can compare their relative strength by observing the net effect (Grajek and Kretschmer 2009).

How did Facebook's rule change affect the role of installed base on usage intensity? As we are capturing the net effect of user composition and network effects, we assess the effect on both individual forces to make a prediction on how the net effect changes. User composition is unlikely to be affected strongly by the rule change as it affects the supply side, but not the demand side (i.e. the users) of applications. Conversely, the rule change affected the advantage of applications with a large installed base in that notifications simply could not be sent out with the same frequency as before. So the same number of installations would result in less updates being sent out.

We therefore expect the impact of network effects to decrease after Facebook's rule change, leaving the effect of user heterogeneity as the dominant effect.

4.2.4 Application age

Application age, i.e. the time since which the application has been launched, also drives usage intensity. Older applications are expected to be used less intensively as usage follows a fad, i.e. users are only interested in the application for a short time. One goal of Facebook's rule change was to reward more compelling applications. Thus, the change helped applications that retain and keep their users engaged.

We therefore expect usage intensity to decay less rapidly with time after the rule change.

4.2.5 Portfolio effects

Portfolio effects matter in cultural industries in which artists or developers create several products. For example, Hendricks and Sorensen (2009) find for the music industry that spillovers between albums exist. Similarly, on Facebook most developers have a portfolio of (often functionally similar) applications. This may lead to users splitting their time across different applications, or to users developing a taste for a specific developer's applications. The net effect of belonging to a large portfolio is therefore an empirical matter. It is interesting to speculate how Facebook's rule change may have changed the role of belonging to such a portfolio. Post-intervention, "new" applications had to attract users by "earning" credit through successful past notifications. As younger applications had less of a history to fall back on, alternative promotion channels mattered more both in attracting new users and in keeping existing users' interest. The alternative channel was cross-promotion through developers' application portfolios.

Therefore, we expect a large portfolio to have a more positive effect following the rule change.

5. Empirics

5.1 Data

We use a unique dataset from Facebook's public directory of applications which included all applications available on the Facebook platform.[15] All application-specific "about"-pages in this directory have been crawled and parsed daily to extract the variables described below.

Even though our data covers the period of September 1, 2007 to June 30, 2008, we focus on a period of twenty weeks around the rule change we study. Our observation period falls in the early phase[16] of the application platform and was characterized by strong growth in terms of users of Facebook's service as well as the number of applications and their users. The number of applications on the platform grew immensely from around 2,000 in early September 2007 to over 18,000 in early 2008.

We obtained records for 18,552 applications, of which 7,784 were active and contained all variables in the 20-week time window around the rule change. The records include data on an application's entry to the platform, its usage by Facebook members, its developer and finally an assignment to certain categories. Further, we computed a number of measures by observing changes made to the directory page as well as from clustering applications by developer name.

[15] Facebook's application directory has been suspended in July 2011 (as announced on https://developers.facebook.com/blog/post/523/, accessed November 3rd 2011). See Figure B.1 for an example of an application page in the directory.

[16] Facebook's platform for applications was launched on May 24, 2007.

Our data is particularly well-suited for analyzing usage intensity for several reasons. First, we have precise measures for usage and usage intensity. Facebook continuously reports how many users have interacted with the application within the previous 24 hours. It also specifies the percentage of all users, i.e. the ratio of active users (last 24 hours) to all users who have installed the application at any given time. Hence, we observe both an application's installed base and its usage intensity. Second, the measures of usage directly indicate the potential for economic success. Third, our data mitigates selection problems originating from deterred entry and observed survival. Developer entry to the platform is frequent due to low entry barriers. More importantly, however, entry costs can be assumed to be homogeneous. Finally, the dataset includes applications that were successful and applications that never reached a meaningful user base. Since data is recorded from the first day an application appears in the directory, information is available independent of the application's following success. This is rather unique particularly for studies on Internet-based industries where determining entry accurately is often difficult due to poor documentation of the early history of a category or firm. Published accounts on the entities often do not appear before they reach a "threshold scale" (Eisenmann 2006, p. 1193).

5.2 Variables

Variable definitions and summary statistics are given in Table 1 and Table 2 gives pairwise correlations.

Table 1: Variable definitions and summary statistics

Variable	Definition	Mean	SD	Min	Max
$UsageIntensity_{it}$	Usage intensity of an application measured as percentage of daily active users of $NumUsers_{it}$	5.07	8.65	0.14	100
$RuleChange_t$	Dummy for the rule change (zero before sixth week of 2008 and one thereafter)	0.58	0.49	0	1
$AppRating_i$	Time-invariant average user rating of an application	3.59	1.38	0.65	5
$UpdateFreq_{it}$	Total number of updates of an application divided by $WeeksActive_{it}$	0.15	0.24	0	5
$NumUsers_{it}$	Number of users that have installed an application (in million)	0.16	1.08	0	69.86
$WeeksActive_{it}$	Weeks since an application has first appeared in Facebook's application directory	13.97	10.38	0	48
$NumSisterApps_{it}$	Number of sister applications offered by the same developer	17.64	51.86	0	333

Notes: The number of observations for all variables is 109,233. All observations are restricted to be within ten weeks of the rule change. Summary statistics are presented in linear form for all variables. In the regressions, the logarithm of $UsageIntensity_{it}$, $NumUsers_{it}$, $WeeksActive_{it}$, and $NumSisterApps_{it}$is used.

Table 2: Pairwise correlations

Variable	[1]	[2]	[3]	[4]	[5]	[6]	[7]
[1] $\ln(UsageIntensity_{it}+1)$	1.000						
[2] $RuleChange_t$	-0.205	1.000					
[3] $AppRating_i$	0.014	0.006	1.000				
[4] $UpdateFreq_{it}$	0.171	0.105	0.027	1.000			
[5] $\ln(NumUsers_{it})$	-0.211	-0.002	-0.120	0.058	1.000		
[6] $\ln(WeeksActive_{it})$	-0.689	0.169	0.066	-0.138	0.341	1.000	
[7] $\ln(NumSisterApps_{it})$	-0.047	-0.155	0.054	-0.098	0.018	-0.017	1.000

Note: All observations are restricted to be within ten weeks of the rule change.

5.2.1 Dependent variable ($\boldsymbol{UsageIntensity_{it}}$)

Our dependent variable is an application $\boldsymbol{i}$'s usage intensity in week $\boldsymbol{t}$ measured as the average percentage of daily active users. Hence, we observe the percentage of an application's installed base of users that uses the application on a given day and form the weekly average. All other time-dependent variables are also observed on a daily basis and then aggregated up to the weekly level.[17] Given the skewness in $\boldsymbol{UsageIntensity_{it}}$, we use the variable's logarithm for our regressions.

5.2.2 Rule change ($\boldsymbol{RuleChange_t}$)

As discussed in section 3, Facebook changed the rules in how far applications can send out notifications in February 2008. We therefore construct a dummy variable $RuleChange_t$ which takes a value of zero before the change (until the sixth week of 2008) and a value of one thereafter.

5.2.3 Application quality ($\boldsymbol{AppRating_i}$)

Users can rate applications on a five-point Likert scale. The variable $AppRating_i$ captures the average user rating of application i. We can only construct a time-invariant measure of application rating as Facebook's application directory did not report user ratings before February 2008.

5.2.4 Update activity ($\boldsymbol{UpdateFreq_{it}}$)

We observe several events indicating an update of the underlying application: we check if the name or one of the descriptive elements of an application has changed, if the screenshot was updated, and if the application category has changed. For each point in time, we calculate an application's update intensity $UpdateFreq_{it}$ as the cumulative number of updates divided by application age ($WeeksActive_{it}$).

5.2.5 Installed base ($\boldsymbol{NumUsers_{it}}$)

At the application level, we observe the number of Facebook users that have installed an application on their profile page. As the number of installations of an application is highly skewed, we use the natural logarithm of $NumUsers_{it}$.

5.2.6 Application age ($\boldsymbol{WeeksActive_{it}}$)

We measure the age of an application as the weeks since the application has appeared in the application directory for the first time. We use the natural logarithm of $WeeksActive_{it}$ to allow for a non-linear decay in usage intensity over time.

[17] We aggregate data from daily to weekly level to average out weekday-dependent differences in usage intensity.

5.2.7 Firm portfolio ($NumSisterApps_{it}$)

As developers can release several applications, we measure a firm's portfolio in the form of sister applications. $NumSisterApps_{it}$ is the number of an application i's sister applications at time t.

5.3 Estimation strategy

We proceed in two steps. We first run explorative random-effects regressions to assess how the direct influence of an application's perceived quality changes with the rule change. We then use a more conservative set of fixed-effect regressions (controlling for time-invariant application-specific drivers of usage intensity) to test how the rule change moderated drivers of usage intensity.

As seen in section 3, the rule change led to an increase in average application quality. Based on this, we first explore if incentives for providing higher quality have been increased by the rule change, i.e. does the same level of quality result in higher usage intensity? To address this question, we estimate the following baseline specification separately in the periods before and after the rule change:

$$\begin{aligned} UsageIntensity_{it} &= \beta_1 AppRating_{it} + \beta_2 UpdateFreq_{it} + \beta_3 NumUsers_{it} \\ &\quad + \beta_4 WeeksActive_{it} \\ &\quad + \beta_5 NumSisterApps_{it} + Category_{i,1} + Category_{i,2} + u_{it} \end{aligned}$$

In this regression, $UsageIntensity_{it}$ is regressed on $AppRating_{it}$, other drivers of usage intensity, as well as a set of dummies representing an application's categories.[18] Comparing effect sizes of $AppRating_{it}$ before and after the rule change lets us assess if and how the importance of an application's perceived quality changed with the rule change.

Clearly, this random-effect estimation strategy poses challenges to identification as many application-specific characteristics are not controlled for. For example, even if two applications belong to the same (broad) category "just for fun", they can induce different levels of involvement, leading to differences in usage intensity. We therefore treat the results from this model as exploratory evidence only.

To tackle this drawback and to assess how the drivers of usage intensity are moderated by the rule change, we run a more restrictive set of fixed-effect regressions. In these models, application-specific effects capture all time-invariant characteristics that might drive usage intensity. As our measure for application quality is time-invariant, it is absorbed by the application-specific fixed-effects and can no longer be identified directly.[19] We thus run the following regression:

$$\begin{aligned} UsageIntensity_{it} = a_i &+ \beta_1 RuleChange_t + \beta_2 RuleChange_t * AppRating_{it} \\ &+ \beta_3 UpdateFreq_{it} + \beta_4 RuleChange_t * UpdateFreq_{it} \\ &+ \beta_5 NumUsers_{it} + \beta_6 RuleChange_t * NumUsers_{it} \\ &+ \beta_7 WeeksActive_{it} + \beta_8 RuleChange_t * WeeksActive_{it} \\ &+ \beta_9 NumSisterApps_{it} + \beta_{10} RuleChange_t * NumSisterApps_{it} + u_{it} \end{aligned}$$

[18] Developers can choose up to two out of 22 category labels for each application. Example categories are "just for fun", "photo", "chat", "dating", or "gaming".

[19] The time-invariant category assignment is also absorbed by the fixed effects.

All time-invariant, application-specific heterogeneity is absorbed by the application fixed-effects a_i. We then include all identified drivers of usage intensity as main effects as well as in interaction terms with the rule change. As discussed, the main effect of application quality is absorbed by the fixed effects. However, the interaction term with the rule change can still be identified and allows answering if application quality became a more important driver for usage intensity after the rule change. The other interaction terms also allow identifying the moderating effect of the rule change on the relative importance of the drivers of usage intensity.

Another concern relates to other possibly unobserved shocks affecting the platform. If these shocks drive the effectiveness of the identified drivers of usage intensity, the dummy for the rule change could also capture these shocks and not only the actual rule change. To mitigate this possibility as far as possible, we restrict our analysis to a short time period around the rule change.[20]

5.4 Results

The results for the random-effects regressions with cluster-robust standard errors at the application-level are given in Table 3. The sample in column (3-1) is restricted to the ten-week period before the rule change, whereas columns (3-2) and (3-3) are restricted to the ten weeks after the change. In the first and third columns, only applications launched before the rule change are included, while the second column includes newly launched applications. Comparing coefficients in (3-1) and (3-3) shows how the drivers of usage intensity changed for the same set of applications. By contrast, comparing coefficients in (3-1) and (3-2) shows how the drivers of usage intensity for the full set of applications changed.

Table 3: Random-effect model: how does the importance of application quality change?

DEPENDENT VARIABLE: $UsageIntensity_{it}$			
	(3-1)	(3-2)	(3-3)
Applications	Only old	Old+new	Only old
INDEPENDENT VARIABLES	Before Change	After Change	After Change
$AppRating_i$	0.00887*	0.0279***	0.0415***
	(0.00491)	(0.00457)	(0.00493)
$UpdateFreq_{it}$	0.188***	0.166***	0.187***
	(0.0211)	(0.0217)	(0.0349)
$NumUsers_{it}$	-0.0225***	-0.115***	-0.117***
	(0.00368)	(0.00324)	(0.00329)
$WeeksActive_{it}$	-0.872***	-0.576***	-0.534***
	(0.00729)	(0.00665)	(0.00817)
$NumSisterApps_{it}$	-0.0307***	-0.0309***	-0.0202***
	(0.00266)	(0.00454)	(0.00452)
Observations	46,312	62,921	51,882
Number of Applications	6,012	7,595	5,838
R^2	0.633	0.452	0.353

Notes: Random-effect OLS point estimates with standard errors clustered on the application-level in parentheses. Asterisks denote significance levels (*** p<0.01, ** p<0.05, * p<0.1). A

[20] For the main results, we use a time window from ten weeks before the rule change to ten weeks thereafter. In the robustness section, we restrict the sample to five weeks before and after the rule change.

constant and dummies for the application category are estimated but not reported. $UsageIntensity_{it}$, $NumUsers_{it}$, $WeeksActive_{it}$, and $NumSisterApps_{it}$ are in logarithmic form, the other variables enter in linear form. All observations are restricted to be within ten weeks of the rule change.

For the explorative random-effects model, we focus on the change in the direct effect of application quality on usage intensity as well as the main effects of our drivers of usage intensity. The coefficient of $AppRating_{it}$ strongly increases after the rule change, i.e. the same quality level as before the change is rewarded more in terms of usage intensity. The effect of application rating goes up most for the subsample of applications already on the market.[21] In addition to the effect of application quality we also find that developers actively managing their application, i.e. applications with a higher update frequency, experience higher usage. The number of users who have installed an application and the time an application is already on the market are both negative predictors of usage intensity. Finally, a larger pool of sister applications by the same developer reduces usage intensity of the focal application.

We now turn to the fixed-effect model to discuss the moderating effects of the rule change on the drivers of usage intensity. For these regressions, we do not consider the periods before and after the rule change separately but include a timeframe from ten weeks before until ten weeks after the rule change and identify the effects from interactions with the dummy variable $PolicyChange_t$. The first column of Table 4 reports the main effects, while interaction effects are included in the second and third specifications. The difference between the last two specifications is that in (4-2) all applications are considered, while in (4-3) only applications launched before the rule change are included.

The coefficients on the main effects of update frequency, number of installations, weeks active, and number of sister applications (presented in the first column) have the same sign and significance as in the random-effects model. The coefficients also maintain their sign and significance when the interaction terms are added in columns two and three. As discussed, the main effect of the application rating is absorbed by the rule change, but the interaction term with the rule change reveals a clearly increased importance of application rating as the rule change is enacted. The benefits from higher update frequency increase with the rule change, while the number of an application's users becomes an even stronger negative driver of usage intensity. The coefficient for the interaction of rule change with weeks active as well as with the number of sister applications is positive and significant, which suggests that applications stay attractive for longer and focal applications with many sister applications achieve a comparably higher usage intensity. Note that the size of coefficients for the interaction terms is very similar in size between the full sample and the sample of applications launched before the rule change.

[21] The endogeneity of quality may be a concern for applications launched after the rule change. For earlier ones this is not problematic as the application was introduced (and its quality was fixed) prior to the change.

Table 4: Fixed-effect model: how are the drivers of usage intensity affected by the rule change?

DEPENDENT VARIABLE: $UsageIntensity_{it}$	(4-1)	(4-2)	(4-3)
Applications	Old+new	Old+new	Only old
INDEPENDENT VARIABLES	Baseline Regression	Rule Change	Rule Change
$RuleChange_t$	0.0100**	-0.315***	-0.300***
	(0.00490)	(0.0223)	(0.0224)
$AppRating_i * RuleChange_t$		0.0377***	0.0373***
		(0.00303)	(0.00303)
$UpdateFreq_{it}$	0.173***	0.134***	0.160***
	(0.0195)	(0.0201)	(0.0218)
$UpdateFreq_{it} * RuleChange_t$		0.0722***	0.124***
		(0.0246)	(0.0297)
$NumUsers_{it}$	-0.230***	-0.170***	-0.172***
	(0.00595)	(0.00615)	(0.00663)
$NumUsers_{it} * RuleChange_t$		-0.0411***	-0.0419***
		(0.00181)	(0.00181)
$WeeksActive_{it}$	-0.653***	-0.742***	-0.744***
	(0.00689)	(0.00779)	(0.00825)
$WeeksActive_{it} * RuleChange_t$		0.188***	0.184***
		(0.00646)	(0.00638)
$NumSisterApps_{it}$	-0.00488*	-0.0166***	-0.0167***
	(0.00286)	(0.00285)	(0.00287)
$NumSisterApps_{it} * RuleChange_t$		0.0114***	0.0121***
		(0.00255)	(0.00254)
Observations	109,233	109,233	98,194
Number of Applications	7,784	7,784	6,027
R^2	0.692	0.708	0.727

Notes: Fixed-effect OLS point estimates with standard errors clustered on the application-level in parentheses. Asterisks denote significance levels (*** p<0.01, ** p<0.05, * p<0.1). A constant is included but not reported. $UsageIntensity_{it}$, $NumUsers_{it}$, $WeeksActive_{it}$, and $NumSisterApps_{it}$ are in logarithmic form, the other variables enter in linear form. All observations are restricted to be within ten weeks of the rule change.

5.5 Robustness checks

We restricted our analysis to a short timeframe around the rule change to avoid confounding the effects from the rule change with other contemporaneous trends. To test the robustness of our results, we further restrict the observation window to five weeks around the rule change. Table A.1 presents results for the random-effect model while Table A.2 presents results for the fixed-effect model. The results with these more challenging restrictions still hold. The effect of application rating becomes approximately 30% weaker both in the random- and in the fixed-effect models: in the random-effects model the difference between (A1-1) on the one hand and (A1-2) and (A1-3) on the other becomes smaller; in the fixed effect model the coefficient of the interaction term decreases when reducing the sample period.

As an additional robustness check, we use the logarithm of the number of daily active users as an alternative dependent variable. This variable captures total activity for an

application instead of the per-user measure usage intensity and can therefore be interpreted as a proxy for an application's total profit potential instead of per-user profit potential. Regarding sign and significance, the results in Table A.3 are very similar to the results for the dependent variable usage intensity. The only notable difference is that the interaction terms between the number of sister apps and the rule change lose significance. The stability between our result sets suggests that overall usage is predominantly driven by usage intensity rather than the sheer number of users.

6. Discussion and Conclusion

6.1 Interpretation of the results

Our empirical results show that the rule change initiated by Facebook, the platform owner, had a profound impact on the determinants of success in the market for Facebook applications. All drivers of usage intensity, the most meaningful measure of application success, were affected in their impact on usage intensity by Facebook's rule change.

Facebook's move was first and foremost designed to "incentivize developers to improve the quality of their notifications". Given the nature of notifications – a term capturing both invitations for new users (designed to encourage new installations) and activity reports to other existing users (designed to keep engagement high) – notifications are an important part of a user's application experience. The first robust result we find thus confirms Facebook's stated aim of improving application quality. This is visible in Figure 1, where there is a marked jump in quality ratings for applications released after the change, but also reflected in the random-effects regressions which show that quality matters more (both for old and new applications) after the change. Most rigorous econometrically are the results controlling for time-invariant application characteristics in Table 4. Here, while we do not get a coefficient on quality (as it is absorbed by the application fixed-effect), we find that post-change application rating has a stronger positive impact on usage intensity. Thus, the new way in which notifications work rewarded the applications which successfully kept users engaged through compelling notifications.

The frequency of updates positively affects usage intensity. This supports the intuition that actively managed applications (i.e. frequently updated ones) enjoy higher usage intensity after the rule change as user engagement is rewarded more. This is especially relevant for older applications that have been introduced under a regime which did not incentivize high quality greatly. These applications can "catch up" by actively managing quality (and maintaining user interest through updates) after the change.

We then consider the net effect of network effects and user composition. In line with expectations, network effects matter less for usage intensity after the change. There are two related explanations for this. First, it is more difficult for widely used, but not particularly engaging applications to leverage their sheer size to keep users' interest. That is, if users receive plenty of notifications from a particular application, they may eventually start using the application more since the notifications suggest intense activity. Second, notifications are another mechanism of engaging users. In this case, users are not simply triggered to engage through friends using the application, but also through notifications. Thus, the two substitute for each other to some extent, rendering sheer network size relatively less important.

Another direct implication of the increased quality incentives for applications after the change is the intuition that applications "age well" post-change. That is, the (negative) fad effect is less pronounced after Facebook's intervention. It is interesting to note that this is relevant even though we control for other factors like update frequency and quality (and other time-invariant characteristics), so this result indeed suggests that applications decay more

slowly as a consequence of the rule change. "Compelling notifications" are one driver of this, a change in the composition of active applications is another.

Finally, a firm's application portfolio becomes a more important driver of usage intensity after the rule change. Before the change, application developers could rely mostly on the free viral channels provided by Facebook to distribute their applications and to keep existing users engaged. In this phase, the advertising space developers had available in each application was probably mostly used for generating revenues from external advertising networks and not to maintain the own application network. As Facebook reduced the freely available information, an own "advertising network" of Facebook applications became an increasingly valuable resource for a developer. Developers with a network of multiple applications can tap into this resource to maintain usage intensity across their network.

Our results confirm widespread changes in the success factors and consequently the market structure of the market for Facebook applications. We do not know the precise goals Facebook had in mind when initiating this change, but our results are in line with an incentive to reward high-quality applications and drive low-quality ones out of the market. This may also trigger a concentration towards one dominant application in every genre as higher quality is rewarded with further opportunities to capitalize on it. So while keeping entry barriers low helps keeping up the "long tail" (Anderson 2006), bandwagon behavior (Dellarocas et al. 2010; Duan et al. 2009) may also lead to a "superstar effect" (Rosen 1981).

One can only surmise that these goals should eventually lead to increased monetization opportunities for Facebook, which is supported to some extent by the fact that Facebook later implemented a change concerning the way monetary transactions are channeled through the Facebook website (and that lets Facebook keep a share of revenues).

6.2 Managerial implications

Our results have a number of implications for practice. First, we find that quality can be incentivized through "soft", i.e. non-excluding rules. This is an alternative to the costly (and error-prone) "quality threshold" rule under which the platform owner exerts control over which software appears on the platform through a vetting and quality assessment process. While such an approach may increase the average quality of active applications, it may also be counterproductive in a nascent market in which consumer preferences are not (yet) settled and there may be innovative applications that would fail the established quality criteria. Second, we find that the drivers of application success are contingent on the environment set by the platform owner. This includes the promotion channels available to applications and the opportunity to implement technical (or layout) changes after an application's launch.

6.3 Limitations and further research

Our study is not without limitations and is exploratory in several respects. First, our data is aggregated at the application level. Thus, we can observe changes in the aggregate behavior of users of an application, but not how an individual's behavior changes over time. Especially given that post-change application usage decays less rapidly it would be interesting to gather individual-level data to see what drives this result. Second, we do not observe profits or revenues, neither by applications developers or the platform owner, Facebook. Hence, we cannot infer precisely if the rule change worked in the intended way or if applications developers benefited from this on average. However, we feel it is reasonable to assume that developers' actions reveal their preferences and that the upward shift in quality was a response designed to exploit the new environment. Similarly, Facebook's stated aim of increasing quality and user satisfaction is presumably (and plausibly) related to future (and current) opportunities for monetization for Facebook. Third, we study a particular episode in the evolution of a single, albeit the most important, social networking platform. We should be

careful therefore in extrapolating the results to other platforms and rule changes. Nevertheless, our study offers an appealing look into the various ways in which platform owners can manage their platforms.

In summary, we study the changes in the success factors of Facebook applications following a change in the way applications could send out notifications. Our results suggest that application developers respond to "soft" quality incentives by launching better applications, in line with the goals that Facebook stated when announcing the change in notification rule. This study contributes to the emerging literature on the empirics of platform markets and we hope that it sheds light on the interaction between rules set by the platform owner and market dynamics in complementary goods markets.

References

ABRAHAMSON, E. & ROSENKOPF, L. 1993. Institutional and competitive bandwagons: Using mathematical modeling as a tool to explore innovation diffusion. *Academy of Management Review,* 18**,** 487-517.

AGARWAL, R., ANIMESH, A. & PRASAD, K. 2009. Research Note—Social Interactions and the "Digital Divide": Explaining Variations in Internet Use. *Information Systems Research,* 20**,** 277-294.

ANDERSON, C. 2006. *The long tail: How endless choice is creating unlimited demand,* London, Random House.

ANIMESH, A., VISWANATHAN, S. & AGARWAL, R. 2011. Competing "Creatively" in Sponsored Search Markets: The Effect of Rank, Differentiation Strategy, and Competition on Performance. *Information Systems Research,* 22**,** 153-169.

ARAL, S. & WALKER, D. 2011. Creating Social Contagion Through Viral Product Design: A Randomized Trial of Peer Influence in Networks. *Management Science,* 57**,** 1623-1639.

BAMPO, M., EWING, M. T., MATHER, D. R., STEWART, D. & WALLACE, M. 2008. The Effects of the Social Structure of Digital Networks on Viral Marketing Performance. *Information Systems Research,* 19**,** 273-290.

BANERJEE, A. V. 1992. A simple model of herd behavior. *The Quarterly Journal of Economics,* 107**,** 797.

BOUDREAU, K. 2008. Too Many Complementors? Evidence on Software Firms. Available at SSRN: http://ssrn.com/abstract=943088.

BOUDREAU, K. & HAGIU, A. 2009. Platform Rules: Regulation of an Ecosystem by a Private Actor. *In:* GAWER, A. (ed.) *Platforms, Markets and Innovation.* Cheltenham, UK and Northampton, MA, US: Edward Elgar.

BOUDREAU, K., LACETERA, N. & LAKHANI, K. 2008. Parallel search, incentives and problem type: Revisiting the competition and innovation link. *Harvard Business School Technology & Operations Mgt. Unit Working Papers.*

BOYD, D. & ELLISON, N. 2007. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication,* 13**,** 210-230.

CABRAL, L. 2006. Equilibrium, epidemic and catastrophe: Diffusion of innovations with network effects. *New Frontiers in the Economics of Innovation and New Technology: Essays in Honor of Paul David, Edward Elgar, London***,** 427-437.

CAILLAUD, B. & JULLIEN, B. 2003. Chicken & egg: Competition among intermediation service providers. *RAND Journal of Economics,* 34**,** 309-328.

CASADESUS-MASANELL, R. & HALABURDA, H. 2011. When Does a Platform Create Value by Limiting Choice? *Harvard Business School Working Paper 11-030.*

CHURCH, J. & GANDAL, N. 2000. Systems Competition, Vertical Merger, and Foreclosure. *Journal of Economics and Management Strategy,* 9**,** 25-51.

COUGHLAN, P. J. 2004. The Golden Age of Home Video Games: from the reign of Atari to the rise of Nintendo. *Harvard Business School Case Study 9-704-487.*

DELLAROCAS, C. 2003. The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science,* 49**,** 1407-1424.

DELLAROCAS, C., GAO, G. & NARAYAN, R. 2010. Are Consumers More Likely to Contribute Online Reviews for Hit or Niche Products? *Journal of Management Information Systems,* 27**,** 127-158.

DUAN, W., GU, B. & WHINSTON, A. B. 2009. Informational cascades and software adoption on the internet: an empirical investigation. *MIS Quarterly,* 33**,** 23-48.

EISENMANN, T. 2006. Internet companies growth strategies: Determinants of investment intensity and long-term performance. *Strategic Management Journal,* 27**,** 1183-1204.

EVANS, D. 2003. Some empirical aspects of multi-sided platform industries. *Review of Network Economics,* 2**,** 191-209.

EVANS, D. 2008. The economics of the online advertising industry. *Review of Network Economics,* 7**,** 359-391.

FARRELL, J. & KLEMPERER, P. 2007. Coordination and Lock-In: Competition with Switching Costs and Network Effects. *In:* ARMSTRONG, M. & PORTER, R. (eds.) *Handbook of Industrial Organization.* Elsevier.

GNYAWALI, D. R., FAN, W. & PENNER, J. 2010. Competitive Actions and Dynamics in the Digital Age: An Empirical Investigation of Social Networking Firms. *Information Systems Research,* 21**,** 594-613.

GODES, D. & MAYZLIN, D. 2009. Firm-created word-of-mouth communication: Evidence from a field test. *Marketing Science,* 28**,** 721-739.

GODES, D., MAYZLIN, D., CHEN, Y., DAS, S., DELLAROCAS, C., PFEIFFER, B., LIBAI, B., SEN, S., SHI, M. & VERLEGH, P. 2005. The firm's management of social interactions. *Marketing Letters,* 16**,** 415-428.

GOLDFARB, A. & TUCKER, C. E. 2011. Privacy Regulation and Online Advertising. *Management Science,* 57**,** 57-71.

GRAJEK, M. & KRETSCHMER, T. 2009. Usage and diffusion of cellular telephony, 1998–2004. *International Journal of Industrial Organization,* 27**,** 238-249.

HAGIU, A. 2011. Quantity vs. Quality: Exclusion By Platforms With Network Effects. *Harvard Business School Working Paper 11-125.*

HENDRICKS, K. & SORENSEN, A. 2009. Information and the skewness of music sales. *Journal of Political Economy,* 117**,** 324-369.

HERVAS-DRANE, A. 2010. Word of mouth and taste matching: A theory of the long tail. *NET Institute Working Paper No. 07-41*

KATZ, E. & LAZARSFELD, P. 1955. *Personal influence: The part played by people in the flow of mass communications,* Glencoe, IL, Free Press.

KATZ, M. & SHAPIRO, C. 1985. Network Externalities, Competition, and Compatibility. *American Economic Review,* 75**,** 424-440.

KATZ, M. & SHAPIRO, C. 1986. Product compatibility choice in a market with technological progress. *Oxford Economic Papers,* 38**,** 146–65.

OESTREICHER-SINGER, G. & SUNDARARAJAN, A. 2006. Network structure and the long tail of electronic commerce. *Proceedings of the Tenth INFORMS Conference on Information Systems and Technology (CIST 2006).*

PHELPS, J., LEWIS, R., MOBILIO, L., PERRY, D. & RAMAN, N. 2005. Viral marketing or electronic word-of-mouth advertising: Examining consumer responses and motivations to pass along email. *Journal of Advertising Research,* 44**,** 333-348.

ROCHET, J. & TIROLE, J. 2003. Platform competition in two-sided markets. *Journal of the European Economic Association,* 1**,** 990-1029.

ROSEN, S. 1981. The economics of superstars. *American Economic Review,* 71**,** 845-858.

SHAPIRO, C. & VARIAN, H. 1998. *Information rules: a strategic guide to the network economy,* Boston, MA, Harvard Business School Press.

TIWANA, A., KONSYNSKI, B. & BUSH, A. A. 2010. Platform Evolution: Coevolution of Platform Architecture, Governance, and Environmental Dynamics. *Information Systems Research,* 21**,** 675-687.

TUCKER, C. & ZHANG, J. 2011. How Does Popularity Information Affect Choices? A Field Experiment. *Management Science,* forthcoming.

WHEELWRIGHT, S. & CLARK, K. 1992. *Revolutionizing product development: quantum leaps in speed, efficiency, and quality,* New York, Free Press.

Appendix A

Table A. 1: Robustness check: random-effect model within five weeks of rule change

DEPENDENT VARIABLE: $UsageIntensity_{it}$			
	(A1-1)	(A1-2)	(A1-3)
Applications	Only old	Old+new	Only old
INDEPENDENT VARIABLES	Before Change	After Change	After Change
$AppRating_i$	0.0148***	0.0223***	0.0354***
	(0.00519)	(0.00484)	(0.00490)
$UpdateFreq_{it}$	0.172***	0.194***	0.264***
	(0.0245)	(0.0255)	(0.0395)
$NumUsers_{it}$	-0.00608*	-0.122***	-0.113***
	(0.00350)	(0.00309)	(0.00308)
$WeeksActive_{it}$	-0.864***	-0.526***	-0.516***
	(0.00876)	(0.00813)	(0.0105)
$NumSisterApps_{it}$	-0.0196***	-0.0345***	-0.0307***
	(0.00267)	(0.00435)	(0.00438)
Observations	26,283	30,956	27,351
Number of Applications	5,956	6,995	5,821
R^2	0.538	0.367	0.368

Notes: Random-effect OLS point estimates with standard errors clustered on the application-level in parentheses. Asterisks denote significance levels (*** p<0.01, ** p<0.05, * p<0.1).). A constant and dummies for the application category are estimated but not reported. $UsageIntensity_{it}$, $NumUsers_{it}$, $WeeksActive_{it}$, and $NumSisterApps_{it}$ are in logarithmic form, the other variables enter in linear form. All observations are restricted to be within five weeks of the rule change.

Table A. 2: Robustness check: fixed-effect model within five weeks of rule change

DEPENDENT VARIABLE: $UsageIntensity_{it}$	(A2-1)	(A2-2)	(A2-3)
Applications	Old+new	Old+new	Only old
INDEPENDENT VARIABLES	Baseline Regression	Rule Change	Rule Change
$RuleChange_t$	-0.00830*	-0.127***	-0.0201
	(0.00449)	(0.0242)	(0.0234)
$AppRating_i * RuleChange_t$		0.0275***	0.0279***
		(0.00297)	(0.00295)
$UpdateFreq_{it}$	0.208***	0.141***	0.171***
	(0.0283)	(0.0277)	(0.0306)
$UpdateFreq_{it} * RuleChange_t$		0.138***	0.169***
		(0.0285)	(0.0313)
$NumUsers_{it}$	-0.302***	-0.250***	-0.240***
	(0.00720)	(0.00754)	(0.00784)
$NumUsers_{it} * RuleChange_t$		-0.0394***	-0.0398***
		(0.00171)	(0.00171)
$WeeksActive_{it}$	-0.682***	-0.724***	-0.778***
	(0.0108)	(0.0111)	(0.0121)
$WeeksActive_{it} * RuleChange_t$		0.115***	0.0823***
		(0.00680)	(0.00623)
$NumSisterApps_{it}$	-0.00748**	-0.0180***	-0.0201***
	(0.00314)	(0.00308)	(0.00307)
$NumSisterApps_{it} * RuleChange_t$		0.0213***	0.0227***
		(0.00246)	(0.00246)
Observations	57,239	57,239	53,634
Number of Applications	7,158	7,158	5,984
R^2	0.654	0.668	0.685

Notes: Fixed-effect OLS point estimates with standard errors clustered on the application-level in parentheses. Asterisks denote significance levels (*** p<0.01, ** p<0.05, * p<0.1). A constant is included but not reported. A constant is included but not reported. $UsageIntensity_{it}$, $NumUsers_{it}$, $WeeksActive_{it}$, and $NumSisterApps_{it}$ are in logarithmic form, the other variables enter in linear form. All observations are restricted to be within five weeks of the rule change.

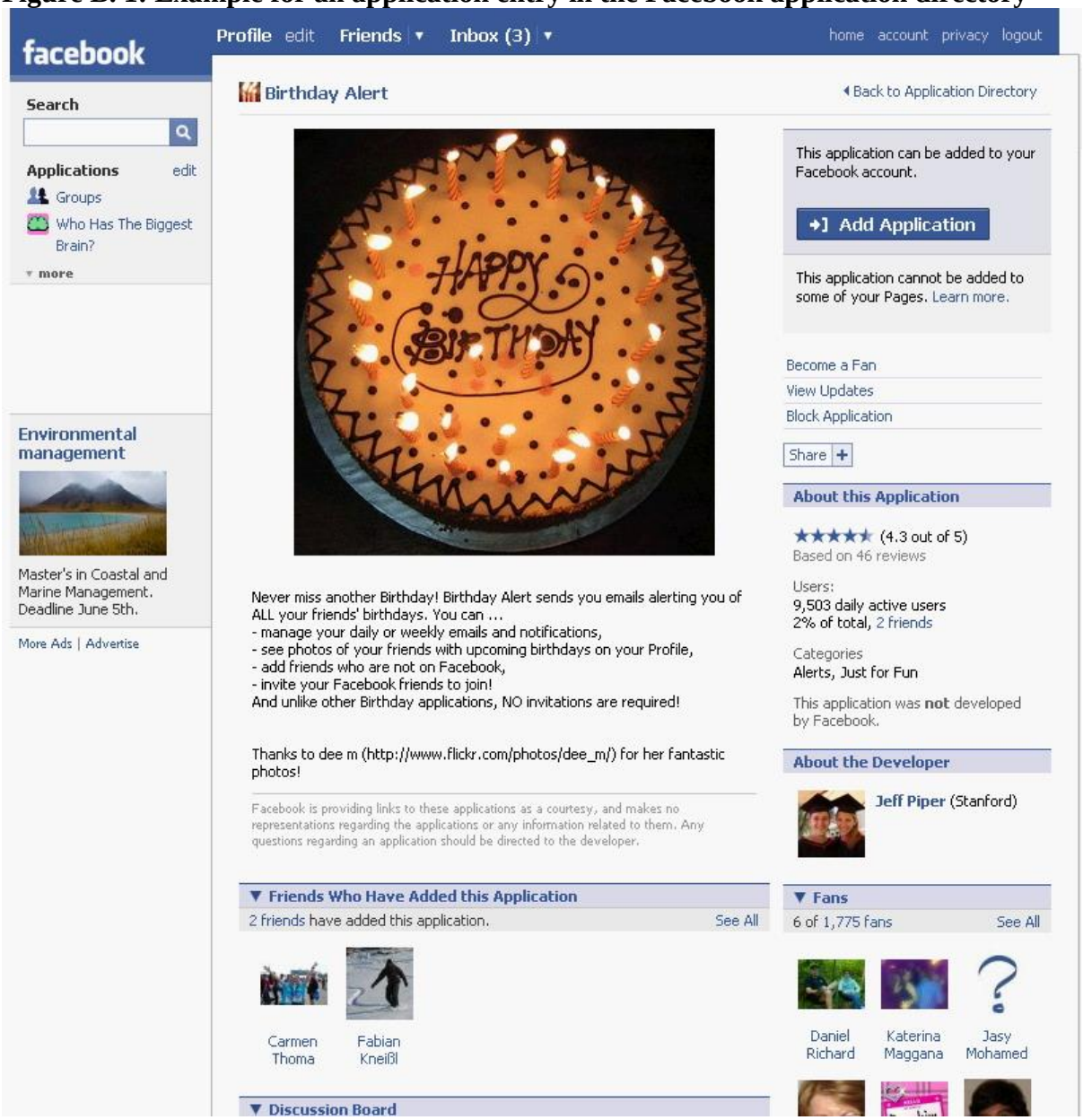
Table A. 3: Robustness check: fixed-effect model with dependent variable number of daily active users

DEPENDENT VARIABLE: $\ln(DailyActiveUsers_{it})$			
	(A2-1)	(A2-2)	(A2-3)
Applications	Old+new	Old+new	Only old
INDEPENDENT VARIABLES	Baseline Regression	Rule Change	Rule Change
$RuleChange_t$	-0.117***	-0.370***	-0.201***
	(0.00717)	(0.0313)	(0.0328)
$AppRating_i * RuleChange_t$		0.0622***	0.0617***
		(0.00461)	(0.00463)
$UpdateFreq_{it}$	0.205***	0.136***	0.191***
	(0.0267)	(0.0285)	(0.0312)
$UpdateFreq_{it} * RuleChange_t$		0.154***	0.264***
		(0.0350)	(0.0430)
$NumUsers_{it}$	0.709***	0.781***	0.764***
	(0.00832)	(0.00867)	(0.00971)
$NumUsers_{it} * RuleChange_t$		-0.0597***	-0.0618***
		(0.00274)	(0.00277)
$WeeksActive_{it}$	-0.774***	-0.874***	-0.890***
	(0.00942)	(0.0108)	(0.0117)
$WeeksActive_{it} * RuleChange_t$		0.188***	0.139***
		(0.00892)	(0.00938)
$NumSisterApps_{it}$	0.00238	-0.0104***	-0.0126***
	(0.00407)	(0.00404)	(0.00409)
$NumSisterApps_{it} * RuleChange_t$		0.00377	0.00461
		(0.00387)	(0.00388)
Observations	109,233	109,233	98,194
Number of Applications	7,784	7,784	6,027
R^2	0.374	0.398	0.393

Notes: Fixed-effect OLS point estimates with standard errors clustered on the application-level in parentheses. Asterisks denote significance levels (*** p<0.01, ** p<0.05, * p<0.1). A constant is included but not reported. A constant is included but not reported. $UsageIntensity_{it}$, $NumUsers_{it}$, $WeeksActive_{it}$, and $NumSisterApps_{it}$ are in logarithmic form, the other variables enter in linear form. All observations are restricted to be within ten weeks of the rule change.

Appendix B (not intended for publication)

Figure B. 1: Example for an application entry in the Facebook application directory



CENTRE FOR ECONOMIC PERFORMANCE
Recent Discussion Papers

1132	Bianca De Paoli Pawel Zabczyk	Cyclical Risk Aversion, Precautionary Saving and Monetary Policy
1131	Carlo Altomonte Filippo De Mauro Gianmarco I. P. Ottaviano Armando Rungi Vincent Vicard	Global Value Chains During the Great Trade Collapse: A Bullwhip Effect?
1130	Swati Dhingra John Morrow	The Impact of Integration on Productivity and Welfare Distortions Under Monopolistic Competition
1129	Gianmarco I. P. Ottaviano	Agglomeration, Trade and Selection
1128	Luis Garicano Claire Lelarge John Van Reenen	Firm Size Distortions and the Productivity Distribution: Evidence from France
1127	Nicholas A. Christakis Jan-Emmanuel De Neve James H. Fowler Bruno S. Frey	Genes, Economics and Happiness
1126	Robert J. B. Goudie Sach Mukherjee Jan-Emmanuel De Neve Andrew J. Oswald Stephen Wu	Happiness as a Driver of Risk-Avoiding Behavior
1125	Zack Cooper Stephen Gibbons Simon Jones Alistair McGuire	Does Competition Improve Public Hospitals' Efficiency? Evidence from a Quasi-Experiment in the English National Health Service
1124	Jörg Claussen Tobias Kretschmer Thomas Spengler	Market Leadership Through Technology - Backward Compatibility in the U.S. Handheld Video Game Industry
1123	Bernardo Guimaraes Kevin D. Sheedy	A Model of Equilibrium Institutions
1122	Francesco Caselli Tom Cunningham Massimo Morelli Inés Moreno de Barreda	Signalling, Incumbency Advantage, and Optimal Reelection Rules
1121	John Van Reenen Linda Yueh	Why Has China Grown So Fast? The Role of International Technology Transfer

1120	Francesco Giavazzi Michael McMahon	The Household Effects of Government Spending
1119	Francesco Giavazzi Michael McMahon	The Household Effects of Government Spending
1118	Luis Araujo Giordano Mion Emanuel Ornelas	Institutions and Export Dynamics
1117	Emanuel Ornelas	Preferential Trade Agreements and the Labor Market
1116	Ghazala Azmat Nagore Iriberri	The Provision of Relative Performance Feedback Information: An Experimental Analysis of Performance and Happiness
1115	Pascal Michaillat	Fiscal Multipliers over the Business Cycle
1114	Dennis Novy	Gravity Redux: Measuring International Trade Costs with Panel Data
1113	Chiara Criscuolo Ralf Martin Henry G. Overman John Van Reenen	The Causal Effects of an Industrial Policy
1112	Alex Bryson Richard Freeman Claudio Lucifora Michele Pellizzari Virginie Perotin	Paying for Performance: Incentive Pay Schemes and Employees' Financial Participation
1111	John Morrow Michael Carter	Left, Right, Left: Income and Political Dynamics in Transition Economies
1110	Javier Ortega Gregory Verdugo	Assimilation in Multilingual Cities
1109	Nicholas Bloom Christos Genakos Rafaella Sadun John Van Reenen	Management Practices Across Firms and Countries
1108	Khristian Behrens Giordano Mion Yasusada Murata Jens Südekum	Spatial Frictions

The Centre for Economic Performance Publications Unit
Tel 020 7955 7673 Fax 020 7955 7595
Email info@cep.lse.ac.uk Web site http://cep.lse.ac.uk

Promotional Reviews: An Empirical Investigation of Online Review Manipulation

Dina Mayzlin, Yaniv Dover, and Judith Chevalier[1]

Yale University

June 2012

[1]The authors contributed equally, and their names are listed in reverse alphabetical order. We thank the seminar participants at the Yale Center for Customer Insights Conference, the Marketing Science Conference in Houston, and the 9th ZEW Conference on the Economics of Information and Communication Technologies, University of Pennsylvania marketing seminar, NYU Information Systems seminar, Dartmouth marketing seminar, Washington University St. Louis marketing seminar, University of Houston marketing seminar, and Stanford marketing seminar. Correspondence: 135 Prospect St., P.O. Box 208200, New Haven, CT 06520, e-mail: dina.mayzlin@yale.edu, yaniv.dover@yale.edu, judith.chevalier@yale.edu.

Abstract

Online reviews have been shown to impact consumer behavior. However, the authenticity of online user reviews remains a concern because, on many sites, firms can manufacture positive reviews for their own products and negative reviews for their rivals. In this paper, we marry the diverse literature on economic subterfuge with the literature on organizational form. We undertake an empirical analysis of promotional reviews, examining both the extent to which fakery occurs and the market conditions that encourage or discourage promotional reviewing activity. Specifically, we examine hotel reviews, exploiting the organizational differences between two travel websites: Expedia.com, and Tripadvisor.com. While anyone can post a review on Tripadvisor, a consumer could only post a review of a hotel on Expedia.com if the consumer actually booked at least one night at the hotel through the website. We examine differences in the distribution of reviews for a given hotel between Tripadvisor and Expedia. We show in a simple model that the net gains from promotional reviewing are likely to be highest for independent hotels that are owned by single-unit owners and lowest for branded chain hotels that are owned by multi-unit owners. Our methodology thus isolates hotels with a disproportionate incentive to engage in promotional reviewing activity. We show that hotels with a high incentive to fake have a greater share of five star (positive) reviews on Tripadvisor relative to Expedia. Furthermore, we show that the hotel neighbors of hotels with a high incentive to fake have more one star (negative) reviews on Tripadvisor relative to Expedia.

PRELIMINARY AND INCOMPLETE- DO NOT CITE WITHOUT PERMISSION.

1 Introduction

User-generated online reviews have become an important resource for consumers making purchase decisions; an extensive and growing literature documents the influence of online user reviews on the quantity and price of transactions.[1] In theory, online reviews should create producer and consumer surplus by improving the quality of the match between consumers and products. However, one important impediment to the improvement in match quality is the possible existence of fake or "promotional" online reviews. Specifically, reviewers with a material interest in the consumer's purchase decision may post reviews that are designed to influence consumers and to resemble the reviews of disinterested consumers. While there is a substantial economic literature on misrepresentation (reviewed below), the specific context of advertising disguised as user reviews has not been extensively studied.

The presence of undetectable (or difficult to detect) fake reviews may have at least two deleterious effects on consumer and producer surplus. First, consumers who are fooled by the promotional reviews may make suboptimal choices. Second, the presence or potential presence of biased reviews may lead consumers to mistrust reviews. This in turn forces consumers to disregard or underweight helpful information posted by disinterested reviewers. For these reasons, the Federal Trade Commission in the United States recently updated its guidelines governing endorsements and testimonials to also include online reviews. According to the guidelines a user must disclose the existence of a material connection between himself and the manufacturer.[2] To the best of our knowledge, there has not been wide-scale enforcement of these laws in the United States, although the FTC did investigate (but did not fine) Ann Taylor LOFT for breaking the law in giving bloggers gift cards for coverage of its fashion show (see Zmuda (2010)). Relatedly, in February 2012, the UK Advertising Standards Authority ruled that TripAdvisor must not claim that it offers "honest, real, or trusted" reviews from "real travelers". The Advertising Standards Authority, in its decision, held that TripAdvisor's claims implied that "consumers could be assured that all review content on the

[1] Much of the earliest work focused on the effect of Ebay reputation feedback scores on prices and quantity sold; for example, Resnick and Zeckhauser (2002), Melnik and Alm (2002)), and Resnick et al. (2006). Later work examined the role of consumer reviews on product purchases online; for example, Chevalier and Mayzlin (2006), Anderson and Magruder (2012)), Berger et al. (2010), Chintagunta et al. (2010).

[2] The guidelines provide the following example, "An online message board designated for discussions of new music download technology is frequented by MP3 player enthusiasts...Unbeknownst to the message board community, an employee of a leading playback device manufacturer has been posting messages on the discussion board promoting the manufacturer's product. Knowledge of this poster's employment likely would affect the weight or credibility of her endorsement. Therefore, the poster should clearly and conspicuously disclose her relationship to the manufacturer to members and readers of the message board" (http://www.ftc.gov/os/2009/10/091005endorsementguidesfnnotice.pdf)

TripAdvisor site was genuine, and when we understood that might not be the case, we concluded that the claims were misleading." (www.asa.org/ASA-action/Adjudications).

In order to examine the potential importance of these issues, we undertake an empirical analysis of the extent to which promotional reviewing activity occurs and the firm characteristics and market conditions that result in an increase or decrease in promotional reviewing activity. The first challenge to any such exercise is that detecting promotional reviews is difficult. After all, promotional reviews are designed to mimic unbiased reviews. For example, inferring that a review is fake because it conveys an extreme opinion is flawed; presumably, individuals who had an extremely positive or negative experience with a product may be particularly inclined to post reviews. In this paper, we empirically exploit a key difference in website business models. In particular, some websites accept reviews from anyone who chooses to post a review while other websites only allow reviews to be posted by consumers who have actually purchased a product through the website (or treat "unverified" reviews differently from those posted by verified buyers). If posting a review requires making an actual purchase, the cost of posting disingenuous reviews is greatly increased. We examine differences in the distribution of reviews for a given product between websites where faking is difficult and websites where faking is easy.

Specifically, in this paper, we examine hotel reviews, exploiting the organizational differences between Expedia.com, and Tripadvisor.com. Tripadvisor is a popular website that collects and publishes consumer reviews of hotels, restaurants, attractions and other travel-related services. Anyone can post a review on Tripadvisor. Expedia.com is a website through which travel is booked; consumers are also encouraged to post reviews on the site, but, a consumer can only post a review if the consumer actually booked at least one night at the hotel through the website in the six months prior to the review post. Thus, the cost of posting a fake review on Expedia.com is quite high relative to the cost of posting a fake review on Tripadvisor. Further, since the reviewer had to undertake a credit card transaction on Expedia.com, the reviewer is not anonymous to the website host and thus, the potential for detection might also be higher. [3]

We present a simple analytical model that examines the equilibrium levels of manipulation of two horizontally-differentiated competitors who are trying to convince a consumer to purchase their product. The model demonstrates that the amount of potential reputational risk determines the

[3] As discussed above, TripAdvisor has been criticized for not managing the fraudulent reviewing problem. TripAdvisor recently announced the appointment of a new Director of Content Integrity. Even in the presence of substantial content verification activity on TripAdvisor's part, our study design takes as a starting point the potential for fraud in TripAdvisor's business model relative to Expedia.

amount of manipulation in equilibrium. We marry the insights from this model to the literature on organizational form and organizational incentive structures. Based on the modelas well as on the previous literature we examine the following hypotheses: 1) independent hotels are more likely to engage in review manipulation (post more fake positive reviews for themselves and more fake negative reviews for their next-door competitors) than branded chain hotels, 2) small owners are more likely to engage in review manipulation than large owner hotels, and 3) hotels with a small management company are more likely to engage in review manipulation than hotels that use a large management company.

Our main empirical analysis is akin to a differences in differences approach (although, unconventionally, neither of the differences is in the time dimension). Specifically, we examine differences in the reviews posted at Tripadvisor and Expedia for different types of hotels. For example, consider calculating for each hotel at each website the ratio of five star (the highest) reviews to total reviews. We ask whether the difference in this ratio for Tripadvisor vs. Expedia is higher for independent vs. branded chain hotels, whether the difference is higher for hotels that are owned by large owners vs. small owners, and whether the difference is higher for hotels that use large management companies vs. small management companies. Either difference alone would be problematic. Tripadvisor and Expedia reviews could differ due to differing populations at the site. Independent versus chain hotels could have different distributions of true quality, for example. However, our approach isolates whether the two hotel types' reviewing patterns are significantly different across the two sites. Similarly, we examine the ratio of one star (the lowest) reviews to total reviews for hotels that are close geographic neighbors of independent vs. chain hotels, hotels with small owners vs. large owners, and hotels with large management companies versus small management companies. That is, we measure whether the neighbor of independent hotels fare worse on Tripadvisor than on Expedia, for example.

The results are largely consistent with our hypotheses. That is, we find that hotel characteristics (such as ownership, affiliation and management structure) affect the amount of review manipulation. We find that there is relatively more positive manipulation than negative manipulation, even though the order of magnitude of the two is similar. We also find that the total amount of review manipulation, while economically significant, is relatively modest: we estimate that an independent hotel owned by a small owner will generate 7 more fake positive reviews (out of 114) and 4 more fake negative reviews than a chain hotel with a large owner.

The paper proceeds as follows. In Section 2 we discuss the previous literature. In Section 3 we present a simple analytical model and hypotheses. In Section 4 we describe the data and present summary statistics. In Section 5 we present our methodology and results, which includes main results as well as robustness checks. In Section 6 we conclude and also discuss limitations of the paper.

2 Previous Literature

Broadly speaking, our paper is informed by the literature on firm's strategic communication, which includes research on advertising and persuasion. In advertising models the sender is the firm, and the receiver is the consumer who tries to learn about the product's quality before making a purchase decision. In these models the firm signals the quality of its product through the amount of resources invested into advertising (see Nelson (1974), Milgrom and Roberts (1986), Kihlstrom and Riordan (1984), Bagwell and Ramey (1994), Horstmann and Moorthy (2003)) or the advertising content (Anand and Shachar (2009), Anderson and Renault (2006), Mayzlin and Shin (2011)). In models of persuasion, the receiver can influence the receiver's decision by optimally choosing the information structure (Crawford and Sobel (1982) and Chakraborty and Harbaugh (2010) show this in the case where the sender has private information, while Kamenica and Gentzkow (2011) show this result in the case of symmetric information). One common thread between all these papers is that in all of them the sender's identity and incentives are common-knowledge. That is, the receiver knows that the message is coming from a biased party, and hence is able to to take that into account when making her decision. In contrast, in our paper there is uncertainty surrounding the sender's true identity and incentives. That is, the consumer who reads a user review on Tripadvisor does not know if the review was written by an unbiased customer or by a biased source.

The models that are most closely related to the current research are Mayzlin (2006) and Dellarocas (2006). Mayzlin (2006) presents a model of "promotional" chat where competing firms, as well as unbiased informed consumers post messages about product quality online. Consumers are not able to distinguish between unbiased and biased word of mouth, and try to infer product quality based on online word of mouth. Mayzlin (2006) derives conditions under which online reviews are persuasive in equilibrium: online word of mouth influences consumer choice. She also demonstrates that producers of lower quality products will expend more resources on promotional reviews. Compared to a system with no firm manipulation, promotional chat results in welfare loss due to distortions in

consumer choices that arise due to manipulation. The welfare loss from promotional chat is lower the higher the participation by unbiased consumers in online fora. Dellarocas (2006) also examines the same issue. He finds that there exists an equilibrium where the high quality product invests more resources into review manipulation, which implies that promotional chat results in welfare increase for the consumer. Dellarocas (2006) additionally notes that the social cost of online manipulation can be reduced by developing technologies that increase the unit cost of manipulation and that encourage higher participation of honest consumers.

While the literature has not extensively studied biased reviewing, the potential for biased reviews affecting consumer responses to user reviews has been recognized. Perhaps the most intuitive form of biased review is the situation in which a producer posts positive reviews for its own product. In a well-documented incident, in February 2004, an error at Amazon.com's Canadian site caused Amazon to mistakenly reveal book reviewer identities. It was apparent that a number of these reviews were written by the books' own publishers and authors (see Harmon (2004)).[4] Other forms of biased reviews are also possible. For example, rival firms may benefit from posting negative reviews of each other's products. In assessing the potential reward for such activity, it is important to assess whether products are indeed sufficient substitutes to benefit from negative reviewing activity. For example, Chevalier and Mayzlin (2006) argue that two books on the same subject may well be complements, rather than substitutes, and thus, it is not at all clear that disingenuous negative reviews for other firm's products would be helpful in the book market. Consistent with this argument, Chevalier and Mayzlin (2006) find that consumer purchasing behavior responds less intensively to positive reviews (which consumers may estimate are frequently fake) than to negative reviews (which consumers may assess to be more frequently unbiased). However, there are certainly other situations in which two products are obviously substitutes; for example, in this paper, we hypothesize that two hotels in the same location are substitutes.[5]

A burgeoning computer science literature has attempted to empirically examine the issue of fakery by creating textual algorithms to detect fakery. Since the entire goal of a fake reviewer is to mimic a real reviewer; identifying textual markers of fakery is difficult. For example, the popular

[4]Similarly, in 2009 in New York, the cosmetic surgery company Lifestyle Lift agreed to pay $300,000 to settle claims regarding fake online reviews about itself. In addition, a web site called fiverr.com which hosts posts by users advertising services for $5 (e.g.: "I will drop off your dry-cleaning for $5") hosts a number of ads by people offering to write positive or negative hotel reviews for $5.

[5]In theory, a similar logic applies to the potential for biased reviews of complementary products (although this possibility has not, to our knowledge, been discussed in the literature). For example, the owner of a breakfast restaurant located next door to a hotel might gain from posting a disingenuous positive review of the hotel.

press has widely cited the methodology described in Ott et al. (2011) in identifying fake reviews. The researchers hired individuals on the Amazon Mechanical Turk site to write persuasive fake hotel reviews. They then analyzed the differences between the fake 5-star reviews and "truthful" 5-star reviews on Tripadvisor to calibrate their psycholinguistic analysis. However, it is possible that the markers of fakery that the researchers identify are not representative of differently-authored fake reviews. For example, the authors find that truthful reviews are more specific about "spatial configurations" than are the fake reviews. However, the authors specifically hired fakers who had not visited the hotel. We can not, of course, infer from this finding that fake reviews on Tripadvisor authored by a hotel employee would in fact be less specific about "spatial configurations" than true reviews. Since we are concerned with fake reviewers with an economic incentive to mimic truthful reviewers, we are skeptical that textual analysis can provide durable mechanisms for detecting fake reviews.[6] Some other examples of papers that use textual analysis to determine review fakery are Jindal and Liu (2007), Hu et al. (2012), and Mukherjee and Glance (2012).

Kornish (2009) uses a different approach to detect review manipulation. She looks for evidence of "double voting" in user reviews. That is, one strategy for review manipulation is to post a fake positive review for one's product and to vote this view as "helpful." That is, Kornish (2009) uses a correlation between review sentiment and usefulness votes as an indicator of manipulation. This approach is vulnerable to the critique that there may be other (innocent) reasons for such correlation, such as confirmatory bias: if most people who visit a product's page are positively inclined towards the product, more positive reviews will be marked as useful since these reviews confirm the initial belief.

Previous literature has not examined the extent to which the design of websites that publish consumer reviews can discourage or encourage manipulation. In this paper, we exploit those differences in design by examining Expedia versus Tripadvisor. The literature also has not empirically tested whether manipulation is more pronounced in empirical settings where it will be more beneficial to the producer. Using data on organizational form, quality, and competition, we examine the relationship between online manipulation and market factors which may increase or decrease the incentive to engage in online manipulation. We will detail our methodology below; however, it is important to understand that our methodology does not rely on identifying any particular review as unbiased (real) or promotional (fake).

[6] One can think of the issue here as being similar to the familiar "arms race" between spammers and spam filters.

Since review manipulation involves rule-breaking (most sites ask reviewers to pledge that they are not incentivised to write the review, and that the review represents an honest opinion of the product), this paper also relates to the economics literature on cheating. The most closely related papers in that stream are Duggan and Levitt (2002), Jacob and Levitt (2003), and Dellavigna and Ferrara (2010). In all three papers the authors do not observe rule-breaking or cheating ("throwing" sumo wrestling matches, teachers cheating on student achievement tests, or companies trading arms in embargoed countries) directly. Instead, the authors infer that rule-breaking occurs indirectly. That is, Duggan and Levitt (2002) document a consistent pattern of outcomes in matches that are important for one of the players, Jacob and Levitt (2003) infer cheating from consistent patterns test answers, and Dellavigna and Ferrara (2010). In all of these papers we see that cheaters respond to incentives. Importantly for our paper, Dellavigna and Ferrara (2010) show that a decrease in reputation costs of illegal trades results in more illegal trading. Our empirical methodology is similar to this previous work. First, we also do not observe review manipulation directly and must infer it from patterns in the data. Second, we hypothesize and show that the rate of manipulation is affected by differences in reputation costs for players in different conditions. The innovation in our work is that by using two different platforms with dramatically different costs of cheating we are able to have a benchmark.

Of course, for review manipulation to make economic sense, online reviews must play a role in consumer decision-making. Substantial previous research establishes that online reviews effect consumer purchase behavior (see, for example, Chevalier and Mayzlin, 2006). There is less evidence specific to the travel context. Vermeulen and Seegers (2009) measure the impact of online hotel reviews on consumer decision-making in an experimental setting with 168 subjects. They show that online reviews increase consumer awareness of lesser-known hotels and positive reviews improve attitudes towards hotels. Similarly, Ye and Gu (2009) use data from a major online travel agency in China to demonstrate a correlation between traveler reviews and online sales.

Finally, our research is related to the literature on ownership incentives. Our research design depends on smaller owner operators having sharper incentives to bear costs to post reviews and on larger hotel entities recognizing the potential for negative spillovers from being caught undertaking fraudulent activities for the entity's other properties. In this sense, our research is related to an extensive literature on differences in incentives between company-owned and franchised units of service industry chains (see, for example, Blair and Lafontaine (2005)). However, our unusually

rich dataset allows us to exploit the fact that ownership patterns in the hotel industry are actually quite complicated. For example, as discussed previously, a hotel can be franchised to a quite large franchisee company; that franchisee company is less incentivized to engage in fraudulent activity than a small franchisee. In our paper, we advance the literature on ownership by utilizing data on these complex ownership structures.

3 A Simple Model and Hypotheses

We propose a very simple and stylized model to fix ideas. The game consists of two competing firms, A and B, and a continuum of consumers. The time line of the game is the following:

1. ***Stage I***: Nature draws the true quality of each firm (q_A and q_B). We assume that the firms' true quality is not observable to any of the game's players.[7] The prior belief on the firm qualities are: $q_A \sim Normal(q_0, \sigma_q^2)$ and $q_B \sim Normal(q_0, \sigma_q^2)$. Here, the two firms a priori are identically distributed, but the model can be easily generalized to the case where the prior means are not equal. Unless otherwise noted, we assume that all other parameters of the model are common knowledge.

2. ***Stage II***: The firms set prices (p_A and p_B), which are observed by all the players.

3. ***Stage III***: Each firm can surreptitiously (and simultaneously) manufacture positive reviews for itself and negative reviews for its competitor. The reviews are posted by a third party platform that does not verify the reviewers' identity. That is, consumers can not differentiate between real and manufactured (or biased) user reviews. We denote by $e_{i,i}$ the effort that firm i invests into positive self-promotion (manufactured positive reviews), and by $e_{i,j}$ the effort that firm i invests into negative reviews for firm j. While the actual firms' efforts are not observed by the consumers, consumers do observe the user ratings for both firms. Hence we can think of the set of user ratings (which consists of real and fake reviews) providing a *signal* to the consumer on the firm's true quality. In particular, the signals arising from user

[7] The case where only firms, but not the consumers, observe each other's true quality yields similar results, but is considerably more complicated.

ratings are the following:

$$s_A = q_A + e_{A,A} - e_{B,A} + \varepsilon_A \tag{1}$$

$$s_B = q_B + e_{B,B} - e_{A,B} + \varepsilon_B \tag{2}$$

That is, the signal generated from user reviews on firm A's quality consists of the true quality (q_A), the positive self-promotion effort by firm A ($e_{A,A}$), the negative effort by its competitor ($e_{B,A}$), as well as a noise term (ε_A) that reflects random shocks experienced by unbiased reviews: $\varepsilon_i \sim Normal(0, \sigma_\varepsilon^2)$. We also assume that the noise terms are independent across firms.

4. We model the manipulation effort as costly to the firm. We can think of this cost as the reputation-related risks associated with this kind of promotion. That is, if the firm is caught doing this kind of activity, it will suffer damage to its reputation, where the damage may differ if the firm is caught doing self-promotion or generating negative review for its competitors. The chance of getting caught is increasing (at an increasing rate) in the intensity of the promotional activity: the cost is convex in the manipulation effort. Hence we assume that $\frac{\partial C(e_{i,i},e_{i,j})}{\partial e_{i,i}} > 0$, $\frac{\partial C(e_{i,i},e_{i,j})}{\partial e_{i,j}} > 0$, $\frac{\partial^2 C(e_{i,i},e_{i,j})}{\partial^2 e_{i,i}} > 0$, and $\frac{\partial^2 C(e_{i,i},e_{i,j})}{\partial^2 e_{i,j}} > 0$. The following assumed simple functional form satisfies these conditions: $C(e_{i,i}, e_{i,j}) = \frac{\delta}{2}(e_{i,i})^2 + \frac{\gamma}{2}(e_{i,j})^2$. Here δ signifies the damage caused to the firm if it caught doing self-promotion, and γ the damage if it is posting negative reviews for its competitor.

5. ***Stage IV***: Finally, the consumer chooses the product that maximized her utility. We assume that the products are horizontally differentiated. We use a simple Hotelling model of differentiation to model consumer choice, where firm A is located at $x = 0$, firm B is located at $x = 1$, and the consumer at location x chooses A if

$$E[q_A|s_A] - tx - p_A \geq E[q_B|s_B] - tx - p_B \tag{3}$$

We assume that consumers are uniformly distributed on the interval $[0, 1]$. Since consumers do not observe the true quality directly, their expected utility from A and B is inferred from the signals generated from user reviews.

We next solve for the firms' optimal actions by backward induction. We start with the consumer's inference in stage 4. After observing the signal s_A and s_B, the consumers' posterior beliefs on the firms' qualities are:

$$E[q_A|s_A] = (1-\mu_s)q_0 + \mu_s(s_A - \hat{e}^*_{A,A} + \hat{e}^*_{B,A}) \tag{4}$$

$$E[q_B|s_B] = (1-\mu_s)q_0 + \mu_s(s_B - \hat{e}^*_{B,B} + \hat{e}^*_{A,B}) \tag{5}$$

where $\mu_s = \frac{\sigma_q^2}{\sigma_\varepsilon^2+\sigma_q^2}$ $(0 < \mu_s < 1)$ is the optimal weight that the consumer puts on the firms' reviews, and $\hat{e}^*_{A,A}$ and $\hat{e}^*_{B,A}$ are the inferred equilibrium effort levels since the consumer does not observe the firms' manipulation activity directly.

Assuming market coverage, the consumer who is indifferent between the two products is located at point $\hat{x}$, where

$$\hat{x} = \frac{1}{2} + \frac{E[q_A|s_A] - E[q_B|s_B] + p_B - p_A}{2t} \tag{6}$$

Hence, the market shares of firms A and B are $\hat{x}$ and $1-\hat{x}$, respectively. This implies the following profit functions for firms A and B, respectively in stage 3:

$$\Pi^*_{A,Stage\,3} = \max_{e_{A,A},e_{A,B}} \left(p_A E_{q_A,q_B,\varepsilon_A,\varepsilon_B}\left[\frac{1}{2} + \frac{E[q_A|s_A] - E[q_B|s_B] + p_B - p_A}{2t}\right] - \delta_A \frac{e^2_{A,A}}{2} - \gamma_A \frac{e^2_{A,B}}{2} \right) \tag{7}$$

$$\Pi^*_{B,Stage\,3} = \max_{e_{B,B},e_{B,A}} \left(p_B E_{q_A,q_B,\varepsilon_A,\varepsilon_B}\left[\frac{1}{2} + \frac{E[q_B|s_B] - E[q_A|s_A] + p_A - p_B}{2t}\right] - \delta_B \frac{e^2_{B,B}}{2} - \gamma_B \frac{e^2_{B,A}}{2} \right) \tag{8}$$

Substituting (4) and (5) into (7) and (8), and taking the expectation, we can re-write the firm's maximization problem as the following:

$$\Pi^*_{A,Stage\,3} = \max_{e_{A,A},e_{A,B}} \left(p_A \left[\frac{1}{2} + \frac{\mu_s(e_{A,A} + e_{A,B} - \hat{e}^*_{A,A} - \hat{e}^*_{A,B} + c_A) + p_B - p_A}{2t} \right] - \delta_A \frac{e^2_{A,A}}{2} - \gamma_A \frac{e^2_{A,B}}{2} \right) \tag{9}$$

$$\Pi^*_{B,Stage\,3} = \max_{e_{B,B},e_{B,A}} \left(p_B \left[\frac{1}{2} - \frac{\mu_s(e_{B,B} + e_{B,A} - \hat{e}^*_{B,B} - \hat{e}^*_{B,A} + c_B) + p_A - p_B}{2t} \right] - \delta_B \frac{e^2_{B,B}}{2} - \gamma_B \frac{e^2_{B,A}}{2} \right) \tag{10}$$

where $c_A = -e_{B,A} - e_{B,B} + \hat{e}^*_{B,A} + \hat{e}^*_{B,B}$ and $c_B = -e_{A,B} - e_{A,A} + \hat{e}^*_{A,B} + \hat{e}^*_{A,A}$. Proposition 1 below summarizes the optimal manipulation levels for the firms as well as a key comparative static result:

Proposition 1. *In stage 3 (after the firms have committed to prices p_A and p_B),*[8] *the optimal promotional levels are the following:*

$$e^*_{A,A} = \frac{p_A \mu_s}{2\delta_A t};\; e^*_{A,B} = \frac{p_A \mu_s}{2\gamma_A t} \tag{11}$$

$$e^*_{B,B} = \frac{p_B \mu_s}{2\delta_B t};\; e^*_{B,A} = \frac{p_B \mu_s}{2\gamma_B t} \tag{12}$$

The Corollary below summarizes several key results that we will use in our empirical analysis:

Corollary 1. *The following results are implied by Proposition 1:*

1) A decrease in the reputational costs of manipulation increases the intensity of this activity: $\frac{\partial e^*_{B,A}}{\partial \gamma_B} > 0$, $\frac{\partial e^*_{B,A}}{\partial \gamma_B} > 0$, $\frac{\partial e^*_{B,A}}{\partial \gamma_B} > 0$, $\frac{\partial e^*_{B,A}}{\partial \gamma_B} > 0$.

2) Firms engage in negative manipulation of reviews of their competitors: $e^*_{A,B} > 0$ and $e^*_{B,A} > 0$, *and this activity increases as the costs of manipulation decrease. Hence, a firm that is located close to a competitor will have more negative reviews than a firm has no close competitors (which will have no fake negative reviews), and the number of fake negative reviews is greater if the competitor has lower costs of manipulation.*

[8] The equilibrium promotional levels here represent a partial equilibrium since they take the prices as given. In the Appendix, we solve for the full equilibrium of the game by endogenizing the prices: solving for the equilibrium prices as function of δ, γ and t. We show that the key comparative static - that the firm decreases the amount of review manipulation as the costs of promotion increase remains true in the full equilibrium as well.

Finally, we turn to the effect that review manipulation has on consumer choice. In the basic model consumer can invert the firm's problem and perfectly discounts the amount of manipulation. That is, in equilibrium, $e^*_{A,A} = \widehat{e}^*_{A,A}$, $e^*_{A,B} = \widehat{e}^*_{A,B}$, $e^*_{B,B} = \widehat{e}^*_{B,B}$, and $e^*_{B,A} = \widehat{e}^*_{B,A}$. Since fake reviews are perfectly discounted, the consumer would make the same choices in the current setting where fake reviews are possible and in one where fake reviews are not possible. Despite the fact that fake reviews do not affect consumer choices in equilibrium, firms prefer to post reviews. That is, if the firm chooses not to engage in manipulation, the consumer who expects fake reviews will think that the firm is terrible.

Next we consider a realistic extension of the model which changes the observability assumption. That is, suppose that the consumer does not observe the costs of each firm but forms an expectation on the costs based on prior beliefs. We believe that this assumption is more realistic for our empirical setting. We can show that this results in an outcome where a firm with lower manipulation cost has a higher share and the firm with higher manipulation cost has a lower share compared to the case where review manipulation is not possible. That is, this Proposition shows that manipulation of reviews may create distortions in choices under imperfect observability.

Proposition 2. *Assume for simplicity that $\delta = \gamma$. Suppose that the consumer does not observe the firms' costs of manipulation. That is, with probability α the firm has high cost of manipulation: $\delta = \delta_H$, and with probability $1-\alpha$ the firm has low cost of manipulation: $\delta = \delta_L$. Consider the case where both types pool on price – consumers can not infer the firm's cost of manipulation from the price. Here $e^*_{L,i,i} = e^*_{L,i,j} = \frac{p_A \mu_s}{2\delta_L t}$, $e^*_{H,i,i} = e^*_{H,i,j} = \frac{p_A \mu_s}{2\delta_H t}$, and $\widehat{e}^*_{i,i} = \widehat{e}^*_{i,j} = \frac{p_A \mu_s}{2(\alpha\delta_L + (1-\alpha)\delta_H)t}$.* Here the consumer under-estimates the amount of manipulation for low-cost type of firm and over-estimates the amount of manipulation for high-cost firm: $e^*_{L,i,i} > \widehat{e}^*_{i,i} > e^*_{H,i,i}$ and $e^*_{L,i,j} > \widehat{e}^*_{i,j} > e^*_{H,i,j}$. This results in a higher share for low-cost firm and a lower share for high-cost firm compared to the case with no manipulation.

Based on the results of this simple model, we formulate the following hypotheses:

1. **Hypothesis 1**: A firm with lower potential reputational costs associated with review manipulation will create more fake reviews.

2. **Hypothesis 2**: A firm that is located close to a competitor will have more fake negative reviews than a firm with no close neighbors.

3. **Hypothesis 3**: A firm that is located close to competitor with low potential reputational costs will have more fake negative reviews than a firm that is located next to a competitor with high costs.

4 Data

User generated internet content has been particularly important in the travel sector. In particular, TripAdvisor-branded websites have more than 50 million unique monthly visitors and contain over 60 million reviews. While our study uses the US site, TripAdvisor branded sites operate in 30 countries. As Scott and Orlikowski (2012) point out, by comparison, the travel publisher Frommer's sells about 2.5 million travel guidebooks each year.

Our data derive from multiple sources. First, we identified the 25th to 75th largest US cities (by population) to include in our sample. Our goal was to use cities that were large enough to "fit" many hotels, but not so large and dense that competition patterns among the hotels would be difficult to determine. We then "scraped" data on all hotels in these cities from Tripadvisor and Expedia. Some hotels will not be listed on one or the other site and some hotels will not have reviews on one or the other site (typically, Expedia). At each site, we obtained the text and star values of all user reviews, the identity of the reviewer (as displayed by the site), and the date of the review. We also obtain data from STR, a market research firm that provides data to the hotel industry (www.str.com). To match the data from STR to our Expedia and Tripadvisor data, we use name and address matching. Our data consist of 3082 hotels matched between Tripadvisor, Expedia, and STR. Our biggest hotel city is Atlanta with 160 properties, and our smallest is Toledo, with 10 properties. Of the 3082 hotels matched across sites, 2931 have reviews on both sites.

Table 1 provides summary statistics for review characteristics, using hotels as the unit of observation, for the set of hotels that have reviews on both sites. Unsurprisingly, given the lack of posting restrictions, there are more reviews on Tripadvisor than on Expedia. On average, our hotels have nearly three times the number of reviews on Tripadvisor as on Expedia. Also, the summary statistics reveal that on average, Tripadvisor reviewers are more critical than Expedia reviews. The average Tripadvisor star rating is 3.50 versus 3.95 for Expedia. Based on these summary statistics, it appears that hotel reviewers are more critical than reviewers in other contexts. For example, numerous studies document that eBay feedback is overwhelmingly positive. Similarly, Chevalier and Mayzlin (2006) report average reviews of 4.14 out of 5 at Amazon and 4.45 at barnesandnoble.com

Table 1: User Reviews at Tripadvisor and Expedia

	Mean	Standard deviation	Minimum	Maximum
Number of Tripadvisor reviews	119.6	172	1	1675
Number of Expedia reviews	42.2	63.2	1	906
Average Tripadvisor star rating	3.52	0.75	1	5
Average Expedia star rating	3.95	0.74	1	5
Share of Tripadvisor 1 star reviews	0.14	0.17	0	1
Share of Expedia 1 star reviews	0.07	0.14	0	1
Share of Tripadvisor 5 star reviews	0.31	0.19	0	1
Share of Expedia 5 star reviews	0.44	0.26	0	1
Total number of hotels	2931			

for a sample of 2387 books.

Review characteristics are similar if we use reviews, rather than hotels as the unit of observation. Our dataset consists of 352,854 TripAdvisor reviews and 123,893 Expedia reviews. Of all reviews, 8.1% of TripAdvisor reviews are 1s and 38.0% of Trip Advisor reviews are 5s. For Expedia, 4.7% of all review are 1s while 48.5% of all reviews are 5s. Note that these numbers differ from the numbers in the table because hotels with more reviews tend to have better reviews. Thus, the average share of all reviews that are 1s is lower than the mean share of 1 star reviews for hotels.

We use the STR categorizations to identify the hotel category (economy, midscale, upper-midscale, upscale, upper upscale and luxury) and we used data from STR on the year that the hotel property was built to construct the hotel age. We also use STR to obtain the hotel location; we assign each hotel a latitude and longitude designator and use these to calculate distances between hotels of various types. Most importantly, we use STR data to construct the various measures of organizational form that we use for each hotel in the data set. A hotel can be an independent,

a franchised unit of a chain, or a company-owned unit of a chain. In general, franchising is the primary organizational form for the largest hotel chains in the US. For example, Choice Hotels, Marriott Hotels, and Starwood Hotels are all made up of more than 99% franchised units. Within the broad category of franchised units, there are a wide variety of organizational forms. STR provides us with information about each hotel's owner. The hotel owner (franchisee) can be an individual owner-operator or a large company. For example, Archon Hospitality owns 41 hotels in our focus cities. In Memphis, for example, Archon owns two Hampton Inns (an economy brand of Hilton), a Hyatt, and a Fairfield Inn (an economy brand of Marriott). Typically, the individual hotel owner (franchisee) is the residual claimant for the hotel's profits, although the franchise contract generally requires the owner to pay a share of revenues to the parent brand. Owners often, though not always, subcontract day to day management of the hotel to a management company. Typically, the management company charges a few of 3 to 5 percent of revenue, although agreements which involve some sharing of gross operating profits have become more common in recent years.[9] In some cases, the parent brand operates a management company. For example, Marriott provides management services for approximately half of the franchisee-owned hotels under the Marriott nameplate. Like owners, management companies can manage multiple hotels under different nameplates. For example, Crossroads Hospitality manages 29 properties in our data set. In Atlanta, they manage a Hyatt, a Residence Inn (Marriott's longer term stay-focused brand), a Doubletree, and a Hampton Inn (both Hilton brands). As discussed above, our model suggests that hotels with a relationship to a large company– either parent brand, or owning entity, and possibly management company – have a higher cost of posting promotional reviews and have a lower potential benefit from posting promotional reviews than do hotels that operate independently of such entities. While a consumer can clearly observe whether a hotel is a member of a branded chain, the ownership and management structure of the hotel are more difficult to infer for the consumer.

In constructing variables, we focus both on the characteristics of the hotel and characteristics of the hotel's neighbors. Table 2 provides summary measures of the hotel's own characteristics. First, we construct a dummy for whether the hotel is an independent or part of a branded chain, using the characterizations reported in STR: 18% of hotels in our sample are independent. The top 5 parent companies of branded chain hotels in our sample are: Marriott, Hilton, Choice Hotels, Intercontinental, and Best Western. Second, we construct a dummy for whether the hotel is owned

[9]See O'Fallon and Rutherford (2010).

Table 2: Hotel Affiliation, Ownership and Management and Structure

Hotel Status	Share of All Hotels With Reviews	Share of Independent Hotels	Share of Chain Affiliated Hotels
Independent	0.17	1.00	0.00
Marriott Corporation Affiliate	0.14	0.00	0.17
Hilton Worldwide Affiliate	0.12	0.00	0.15
Choice Hotels Int'l Affiliate	0.11	0.00	0.13
Intercontinental Hotels Grp Affiliate	0.08	0.00	0.10
Best Western Company Affiliate	0.04	0.00	0.04
Multi-unit owner	0.31	0.16	0.34
Multi-unit management company	0.52	0.35	0.55
Multi-unit owner AND multi-unit management company	0.26	0.12	0.24
Total Hotels in Sample = 2931			

by a multi-unit ownership entity identified by STR. For example, non-independent hotels that are not owned by a franchisee but owned by the parent chain will be characterized as owned by a multi-unit ownership entity, but so will hotels that are owned by a large multi-unit franchisee. Furthermore, while independent hotels do not have a parent brand, they are in some cases operated by large multi-unit owners. In our sample, 15% of independent hotels and 33% of branded chain hotels are owned by a multi-unit owners. As discussed above, these larger groups will be more concerned about the reputational spillovers of being caught undertaking promotional reviewing activity. Third, for some specifications, we will also examine hotels operated by large multi-unit management companies, which is the case for 32% of independent hotels and for 54% of branded chain hotels.

We then characterize the neighbors of the hotels in our data. The summary statistics for these measures are in Table 3. That is, for each hotel in our data, we first construct a dummy variable that takes the value of one if that hotel has a neighbor hotel within 0.5km. As the summary statistics show, 76% of the hotels in our data have a neighbor. We next construct a dummy that takes the value of one if a hotel has a neighbor hotel that is an independent. Obviously, this set of ones is a subset of the previous measure; 31% of the hotels in our data have an independent neighbor. We also construct a dummy for whether the hotel has a neighbor that is owned by a multi-unit owner. Again, the set of hotels that have a one for this measure are a subset of the hotels that

Table 3: Hotel Characteristics of Neighbor Hotels Within 0.5 km Radius

Hotel Status	Share of All Hotels With Reviews	Share of Independent Hotels	Share of Chain Affiliated Hotels
Hotel has a neighbor	0.76	0.72	0.77
Hotel has an independent neighbor	0.31	0.27	0.50
Hotel has a multi-unit owner neighbor	0.49	0.52	0.49
Hotel has a multi-unit management entity neighbor	0.59	0.58	0.59

have a neighbor. However, as discussed above, this set is not a proper subset of hotels that have a non-independent neighbor; some independent hotels are owned by multi-unit owners and many non-independent hotels are franchised to a small owner-operator. In our data 48% of the hotels have a neighbor owned by a multi-unit owner company. For some specifications, we also examine the management structure of neighbor hotels. We construct a variable that takes the value of one if a hotel has a neighbor hotel operated by a multi-unit management entity, which is the case for 58% of hotels in our sample. For our robustness specifications, we construct measures of hotel relatedness. A hotel is totally unrelated to another hotel if it is not a brand of the same parent (so, a Courtyard Marriott and Marriott are related), if it is not owned by the same ownership entity, and if it is not managed by the same management company. We construct a dummy variable that equals one if a hotel has a neighbor that is totally unrelated, which is the case for 54% of the hotels. Again, this variable will equal one for a subset of the hotels that have a neighbor of any sort.

5 Methodology and Results

As Section 4 describes, we collect reviews from two sites, Tripadvisor and Expedia. There is a key difference between these two sites which we utilize in order to help us identify the presence of review manipulation: while anybody can post a review on Tripadvisor, only those users who

purchased the hotel stay on Expedia in the past six months can post a review for the hotel.[10] This implies that it is far less costly for a hotel to post fake reviews on Tripadvisor versus posting fake reviews on Expedia; we expect that there would be far more review manipulation on Tripadvisor than on Expedia. In other words, a comparison of the difference in the distribution of reviews for the same hotel could potentially help us identify the presence of review manipulation. However, we can not infer promotional activity from a straightforward comparison of reviews for hotels overall on Tripadvisor and Expedia since the population of reviewers using Tripadvisor and Expedia may differ; the websites differ in characteristics other than reviewer identity verification.

Here we take a differences in differences approach (although, unconventionally, neither of our differences are in the time dimension): for each hotel, we examine the difference in review distribution across Expedia and Tripadvisor and across different competitive/ownership conditions. We use the results of Section 3 to argue that the incentives to post fake reviews will differ across different competitive/ownership conditions. That is, we hypothesize that hotels with greater incentive to manipulate reviews will post more fake positive reviews for themselves and more fake negative reviews for their hotel neighbors on Tripadvisor, and we expect to see these effects in the difference in the distributions of reviews on Tripadvisor and Expedia.

Consider the estimating equation:

$$\frac{NStarReviews_{ij}^{TA}}{Total\,Reviews_{ij}^{TA}} - \frac{NStarReviews_{ij}^{Exp}}{Total\,Reviews_{ij}^{Exp}} = X_{ij}B_1 + Own_{ij}B_2 + NeighOwn_{ij}B_3 + \sum \gamma_j + \varepsilon_{ij} \tag{13}$$

This specification estimates correlates of the difference between the share of reviews on TA that are N star and the share of reviews on Expedia that are N star for hotel i in city j. Our primary interest will be in the most extreme reviews, 1-star and 5-star. X_{ij} contains controls for hotel characteristics; these hotel characteristics should only matter to the extent that Tripadvisor and Expedia customers value them differentially. Specifically, we include the hotel's "official" star categorization common to Tripadvisor and Expedia, dummies for the six categorizations of hotel type provided by STR (economy, midscale, luxury, etc), and hotel age. Own_{ij} contains the own-

[10] Before a user posts a review on Tripadvisor, she has to click on a box that certifies that she has "no personal or business affiliation with this establishment, and have not been offered any incentive or payment originating the establishment to write this review." In contrast, before a user posts a review on Expedia, she must log in to the site, and Expedia verifies that the user actually purchased the hotel within the required time period.

hotel organizational and ownership characteristics. In our primary specifications, these include the indicator variable for independent and the indicator variable for membership in a large ownership entity. $NeighOwn_{ij}$ contains the variables measuring the presence and characteristics of other hotels within 0.5km. Specifically, we include an indicator variable for the presence of a neighbor hotel, an indicator variable for the presence of an independent neighbor hotel, and an indicator variable for the presence of a neighbor hotel owned by a large ownership entity. The variables γ_j are indicator variables for city fixed effects.

We start by examining the effects of own-hotel organizational and ownership characteristics (Own_{ij}) on the incentive to manipulate reviews. We argue that branded chain hotels have a higher reputational cost of review manipulation compared to independent hotels since if any single chain hotel is caught posting fake reviews, all the hotels in the chain will suffer damage to their reputation. Similarly, we argue that a multi-unit hotel owner has a higher reputational cost of manipulation since all the hotel owner's properties will suffer if the hotel is caught faking reviews. Finally, the same argument can be made for a multi-unit management company. In all of these cases, an entity that is associated with more properties has more to lose from being caught manipulating reviews: the negative reputational spillovers are higher. Hence, using hypothesis 1 from Section 3, we claim that 1) independent hotels have a higher incentive to post fake positive reviews (have a higher share of 5-star reviews on Tripadvisor versus Expedia) than branded chain hotels, 2) small owners have a higher incentive to post fake positive reviews than multi-unit owner hotels, 3) hotels with a small management company have a higher incentive to post fake positive reviews than hotels that use multi-unit management company. Finally, an alternative explanation for independent hotels having a higher share of positive reviews on Tripadvisor is that the Tripadvisor population likes independent hotels more than the Expedia population. While we can not rule out this alternative explanation, the same critique does not apply to the small owner variables since the ownership structure is not easily observable by customers or reviewers. That is, since neither the identity of the ownership entity (e.g.: Crossroads Hospitality) nor how many units it owns is observable to the reviewers, it is unlikely that reviewers on the different sites would exhibit different preferences for hotels that are owned by multi-unit entities versus single-unit entities. Similarly, we can argue that the size of the management company should not affect the relative preference for the hotel across the two sites.

Finally, we turn to the effect of $NeighOwn_{ij}$variables on review manipulation. Using hypothesis

2 from Section 3, we claim a hotel with a next-door neighbor will have more fake negative reviews (have a higher share of 1-star reviews on Tripadvisor than on Expedia) than a hotel with no next-door neighbor. In addition, using hypothesis 3 from Section 3, we claim that the following next-door neighbor characteristics will result in an increase in fake negative reviews: 1) having a neighbor that is independent, 2) having a neighbor that has a small owner, and 3) having a neighbor that is managed by a small management company.

5.1 Main Results

In this Section we present the estimation results of the basic differences in differences approach to identify review manipulation. Table 4 presents the results of the estimation of Equation 13). Heteroskedasticity robust standard errors are used throughout. We first turn to the specification where the dependent variable is the difference in the share of 5-star reviews. That is, the dependent variable is:

$\frac{5StarReviews_{ij}^{TA}}{Total\ Reviews_{ij}^{TA}} - \frac{5StarReviews_{ij}^{Exp}}{Total\ Reviews_{ij}^{Exp}}$,

This is our measure of possible positive review manipulation. Consistent with our hypothesis that independent hotels optimally post more positive fake reviews, we see that independent hotels have 2.8 percentage points higher difference in the share of 5-star reviews across the two sites than branded chain hotels. Since hotels on Tripadvisor have on average a 31% share of 5-star reviews, the magnitude of the effect is large. As we mentioned before, while this result is consistent with manipulation, we can not rule out the possibility that reviewers on Tripadvisor tend to prefer independent hotels over branded chain hotels to a bigger extent than Expedia customers.

More interestingly, it is more difficult to believe that there is a strong disparity across sites in preferences for hotels with multi-unit owners, a hotel characteristic that is virtually unobservable to the consumer. Consistent with our hypothesis that multi-unit owners will find review manipulation more costly, and therefore engage in less review manipulation, we find that hotels that are owned by a multi-unit owner have 3.2 percentage point smaller difference in the share of 5-star reviews across the two sites. This translates to about four fewer 5-star reviews on Tripadvisor if we assume that the share of Expedia reviews stays the same across these two conditions and that the hotel has a total of 114 reviews on Tripadvisor, the site average. While we include neighbor effects in this specification, we do not have strong hypotheses on the effect of neighbor characteristics on the difference in the share of 5-star reviews across the two sites, since there is no apparent incentive

for a neighboring hotel to practice positive manipulation on the focal hotel. Indeed, in the 5-star specification, none of the estimated neighbor effects are significant.

We next consider to the specification where the dependent variable is the difference in the share of 1-star reviews. Our dependent variable is thus:

$\frac{1StarReviews_{ij}^{TA}}{TotalReviews_{ij}^{TA}} - \frac{1StarReviews_{ij}^{Exp}}{TotalReviews_{ij}^{Exp}}$,

This is our measure of negative review manipulation. Unlike the previous specification, here, we do not expect to see any effects of the hotel's organizational structure on its own share of 1 star reviews since a hotel is not expected to negatively manipulate its own ratings. Instead, our hypotheses concern the effects of the presence of neighbor hotels on negative review manipulation. The results are in Column 2 of Table 4. Our coefficient estimates suggest that the presence of any neighbor increases the difference in the 1-star share across the two sites, even though the effect is not significant. However, the presence of an independent hotel within 0.5km results in an increase of 1.8 percentage point in the difference in the share of 1-star reviews across the two sites relative to a non-independent neighbor. Our point estimates imply that having an independent neighbor versus having no neighbor results in a 2.8 percentage point increase in 1 star reviews (0.89 percentage points for having any neighbor plus 1.88 for the neighbor being independent). These are large estimated effects given that the average share of 1-star reviews is 15% for a hotel on Tripadvisor. Again, we hypothesize that multi-unit owners bear a higher cost of review manipulation and thus will engage in less review manipulation. Our results show that having a hotel with a multi-unit owner within 0.5km results in 2 percentage point decrease in the difference in the share of 1-star reviews across the two sites, relative to having a neighbor that is a single-unit owner.

What do the results in Table 4 suggest about the extent of manipulation of reviews overall on an open platform such as Tripadvisor? As we discuss above, the amount of manipulation depends on the exact hotel characteristics. As an example, let's consider the difference in positive manipulation under two extreme cases: a) a branded chain hotel that is owned by a multi-unit owner (the case with the lowest predicted and estimated amount of manipulation) and b) an independent hotel that is owned by a small owner (the case with the greatest predicted and estimated amount of manipulation). Our estimates suggest that, assuming the TripAdvisor average of 114 total reviews, we would expect about 7 more positive reviews in case b versus case a. Similarly, we can perform a comparison for the case of negative manipulation by neighbors. Consider case c) being located next door to a branded chain hotel that is owned by a multi-unit owner and (the case with the

Table 4: Estimation Results of Equation 13

		Difference in share of 5 star reviews	Difference in share of 1 star reviews
X_{ij}	Site rating	-0.0140 **	-0.0095 *
		(0.0067)	(0.0054)
	Hotel age	0.0003	0.0005 ***
		(0.0002)	(0.0001)
	Hotel tier controls?	Yes	Yes
Own_{ij}	Hotel is Independent	0.0280***	0.0113
		(0.0102)	(0.0096)
	Multi-unit owner	-0.0322 ***	-0.0028
		(0.0084)	(0.0047)
$NeighOwn_{ij}$	Has a neighbor	-0.0155	0.0091
		(0.0119)	(0.0104)
	Has independent neighbor	-0.0037	0.0191**
		(0.0097)	(0.0079)
	Has multi-unit owner neighbor	-0.0081	-0.0204 ***
		(0.0095)	(0.0073)
γ_j	City-level fixed effects?	YES	YES
	Num. of observations	2931	2931
	R-squared	0.11	0.09

*** p<0.01, ** p<0.05, * p<0.10
Heteroskedasticity robust standard errors in parentheses. All neighbor effects calculated for 0.5km radius.

smallest predicted and estimated amount of manipulation) and case d) being located next door to an independent hotel that is owned by a small owner (the case with the greatest predicted and estimated amount of manipulation). Our estimates suggest that there would be a total of 4 more fake negative reviews in case d versus case c.

While it appears that the total amount of negative manipulation is lower than the amount of positive manipulation, it is useful to note that, given the overall average star rankings on Tripadvisor is above 3, an incremental 1 star review will change the average stars more than an incremental 5 star reviews. Unfortunately, it is impossible for us, given these data, to measure the effect that these ratings changes will have on sales. While Chevalier and Mayzlin (2006) show that 1 star reviews hurt book sales more than 5 star reviews help book sales, those findings do not necessarily apply to this context. Chevalier and Mayzlin note that two competing books on the same subject may indeed be net complements, rather than net substitutes. Authors and publishers, then, may gain from posting fake positive reviews of their own books, but will not necessarily benefit from posting negative reviews of rivals books. Thus, in the contexts of books, 1 star reviews may be more credible than 5 star reviews. We have seen that, in the case of hotels, where two hotels proximate to each other are clearly substitutes, one cannot infer that a 1 star review should be treated by customers as more credible than a 5 star review.

Finally, note that while it appears that the total amount of manipulation is economically significant in that we would expect it to distort choices, the amount of manipulation is small enough so that it should not destroy the informational value of Tripadvisor reviews. That is, we can speculate that while firms engage in review manipulation, and this sometimes distorts consumer choices, consumers still find reviews informative and persuasive. This is of course consistent with the observed popularity of Tripadvisor.

Our preceding analysis is predicated on the hypothesis that promotional reviewers have an incentive to imitate real reviewers as completely as possible. This is in contrast to the computer science literature, described above, that attempts to find textual markets or fake reviews. Nonetheless, we do separately examine one category of "suspicious" reviews. These are reviews that are posted by one-time contributors to Tripadvisor. The least expensive way for a hotel to generate a user review is to create a fictitious profile on Tripadvisor (which only requires an email address), and following the creation of this profile, to post a review. This is, of course, not the only way that the hotel can create reviews. Another option is for a hotel to pay a user with an existing review history to

post a fake review; another possibility is to create a review history in order to camouflage a fake review. Here, we examine "suspicious" reviews– the review for a hotel is the first and only review that the user ever posted. In our sample, 26% of all Tripadvisor reviews are posted by one-time reviewers. These reviews are more likely to be extreme compared to the entire Tripadvisor sample: 24% of one-time reviews are 1-star versus 15% in the entire Tripadvisor sample, and 39% of one-time reviewers are 5-star versus 31% in the entire Tripadvisor sample. Of course, the extremeness of one-time reviews does not in and of itself suggest that one-time reviews are more likely to be fake; users who otherwise do not make a habit of reviewing may be moved to do so by an an unusual experience with a hotel.

In Table 5 we present the results of the following three specifications. In the first column, we present the results of a specification where the dependent variable is the share of one-time contributor user reviews on Tripadvisor. Thus, our dependent variable is:

$\frac{one-time\ Reviews_{ij}^{TA}}{Total\ Reviews_{ij}^{TA}}$.

This captures the incidence of these suspicious reviews and includes potential positive as well as negative manipulation. Consistent with our earlier results, we find that an independent hotel has a 9 percentage point increase in the share of these reviews, which is a very large effect since the average share of one-time reviews amongst all hotels is 26%. Also consistent with our previous results, our point estimates suggest that a multi-unit owner has 1.6 percentage point decrease in the share of these reviews, and neighboring multi-unit hotel results results in a 1.9% decrease in the share. There is one variable in our specification that does not have the anticipated sign. The presence of any neighbor is negatively associated with "suspicious" reviews; our model would predict that this association would be positive.

The other two specifications in Table 5 address the valence of these reviews. For these specifications, the dependent variable is:

$\frac{one-time\ NStarReviews_{ij}^{TA}}{Total\ Reviews_{ij}^{TA}} - \frac{NStarReviews_{ij}^{Exp}}{Total\ Reviews_{ij}^{Exp}}$.

That is, we look at the difference between the share of N-star "suspicious" reviews on TripAdvisor and the overall share of N-Star reviews on Expedia. Column 2 shows the case where N=5. The effect of hotel independence is positive, as predicted, but not significantly different from zero. Multi-unit owner has a statistically significant 2.4 percentage point lower difference in the share of 5-star reviews across the two sites, which is consistent with our hypotheses and earlier results. The neighbor effects are not statistically significant, as they weren't in the specifications that used all

TA 5-star reviews. Column 3 shows the case where N=1. Here, we find that the presence of an independent hotel next door increases the difference in the share of 1-star reviews across the two sites by a statistically significant 2.1 percentage points, while having a hotel owned by a multi-unit owner next door decreases the difference in the share of 1-star reviews by 2.1 percentage points. The presence of a neighbor has an estimated positive effect, as predicted, but as in our previous specifications, is not statistically significant. Overall, these results confirm our prior results that manipulation of reviews takes place in a way that is consistent with predicted hotel incentives.

5.2 Robustness Checks

In this Section, we undertake a number of further checks that the results are robust to a variety of reasonable specifications. First, we consider additional variables concerning hotel structure. Specifically, we include a variable that equals one if the hotel is managed by a multi-unit management company. As we explain in Section 5, the management company is not residual claimant to hotel profitability the way that the owner is, but nonetheless, obviously has a stake in hotel success. Thus, we expect that a multi-unit management company would have a lower incentive to post fake reviews than a single-unit manager (which in many cases is the owner). We also include the neighbor analog of this variable, a variable that takes the value of one if the hotel has a neighbor that is managed by a multi-unit management company. In the first column in Table 6, we use the share difference in 5 star reviews as the dependent variable. We see that indeed a hotel that is managed by a multi-unit management company has a statistically significant 1.9 percentage point decrease in the difference of the share of 5-star reviews between the two sites which we interpret as a decrease in positive manipulation. Notably, the inclusion of this variable does not alter our previous results; independent hotels have more 5-star reviews on TripAdvisor relative to Expedia and hotels with multi-unit owners have fewer. There are, as before, no significant neighbor effects for 5-star reviews. Column 1 of Table 7 , repeats this same specification for 1-star reviews. Here, as before, we have no predictions for the own hotel characteristics (although we do find here that, with the inclusion of the large management dummy, the large owner dummy becomes statistically significantly different from zero). We do have predictions for neighbor characteristics. As before, we find that having an independent hotel neighbor significantly predicts more one star reviews, and that having a large owner chain neighbor predicts fewer one star reviews (although this effect is now not statistically significant). A large management chain is a negative but not statistically significant predictor of

Table 5: Results for Tripadvisor one-time contributor reviewers

		Share of one-time contributor user reviews	Difference in share of 5 stars reviews	Difference in share of 1 stars reviews
X_{ij}	Site rating	-0.0094	-0.0080	-0.0150 **
		(0.0044)	(0.0078)	(0.0074)
	Hotel quality-tier controls?	YES	YES	YES
	Hotel age	0.0006 ***	0.0003	0.0007 ***
		(0.0001)	(0.0002)	(0.0002)
Own_{ij}	Hotel is Independent dummy	0.0916 ***	0.0128	-0.0137
		(0.0083)	(0.0123)	(0.0121)
	Multi-unit owner	-.0160 ***	-0.0245 **	0.0013
		(0.0054)	(0.0118)	(0.0085)
$NeighOwn_{ij}$	Has a hotel neighbor dummy	-0.0164 *	-0.0124	0.0128
		(0.0084)	(0.0157)	(0.0144)
	Has independent hotel neighbor dummy	0.0023	-0.0005	0.0214 *
		(0.0066)	(0.0128)	(0.0118)
	Has a neighbor that is multi-unit owner dummy	-0.0192 ***	-0.0111	-0.0212 *
		(0.0065)	(0.0130)	(0.0110)
γ_j	City-level fixed effects?	YES	YES	YES
	Num. of observations	3063	2874	2874
	R-squared	0.29	0.06	0.08

*** p<0.01, ** p<0.05, * p<0.10
Heteroskedasticity robust standard errors in parentheses. All neighbor effects calculated for 0.5km radius.

one star reviews. Contrary to our hypothesis, the hotel neighbor dummy becomes negative in this specification, although also not statistically significant. Altogether, there is suggestive evidence that larger management companies are associated with less review manipulation.

In the second column in Table 6 and Table 7, we present specifications that include a new neighbor variable: a dummy that is one if the neighbor has absolutely no affiliation with the focal hotel (either via the same parent brand, the same owner company, or the same management company). We do not find that this variable has a significant effect (although we would not expect it to in the 5-star review specifications). Also, the other coefficients are not significantly impacted by including this variable.

In the right column of Table 6 and Table 7, we present 5-star and 1-star specifications that also include hotel chain fixed effects for the ten largest hotel brands. Inclusion of these chain fixed effects allows TripAdvisor and Expedia patrons to have a very general form of different preferences. They can have not only different preferences for hotel quality tiers and hotel age (all included in the controls in our base specifications), but also can have different preferences for different individual hotel brands. These specifications produce results very similar to the base specifications discussed in Table 5. The only change that inclusion of this variable causes compared to the earlier results is that the independent own hotel dummy in the 5-star specification is no longer statistically significant.

We also examine the relationship between our results and the results that would obtain by substituting data from Expedia for data from Orbitz. Until recently, Orbitz, like Expedia, only accepted reviews from individuals who had booked their stay at orbit.com. Starting in late 2010, Orbitz allowed others to submit hotel reviews, but reviews from verified customers are identified as "Verified" and are given higher weight in calculating the Orbitz Reviewer Score for each property. In our robustness results, we use only verified reviews from Orbitz. Thus, these reviews are analogous to Expedia reviews. Summary statistics are shown in Table 8. Orbitz is less attractive to us as a review site than Expedia. There are 104 hotels that have reviews at TripAdvisor and Expedia but no reviews at Orbitz. For hotels with reviews at both Orbitz and TripAdvisor, the hotels have only about three-quarters the number of Orbitz reviews as Expedia for the hotels in our sample. However, if our results are driven by important (and subtle) differences between the customer pools at Expedia and TripAdvisor, robustness of our results for Orbitz may be valuable. [

We do not use unverified reviews from Orbitz because there are very few of them. For our hotels, we have a total of 87716 verified Orbitz reviews and only 692 unverified reviews. The

Table 6: Specifications with 5-star reviews as dependent variable

		Difference in share of 5-star reviews	Difference in share of 5-star reviews	Difference in share of 5-star reviews
X_{ij}	Site rating	-0.0125 *	-0.0137 **	-0.0137 ***
		(0.0068)	(0.0067)	(0.0068)
	Hotel quality-tier controls?	YES	YES	YES
	Hotel age	0.0003	0.0003	0.00004
		(0.0002)	(0.0002)	(0.0002)
Own_{ij}	Hotel is Independent dummy	0.0255 **	0.0290 **	0.0097
		(0.0102)	(0.0103)	(0.0120)
	Hotel is part of a multi-unit owner dummy	-0.0264 ***	-0.0315 ***	-0.0199 **
		(0.0086)	(0.0083)	(0.0086)
	Hotel is managed by a multi-unit management company dummy	-0.0198 **	–	–
		(0.0091)		
	Hotel chain specific dummy	–	–	YES
$NeighOwn_{ij}$	Has a hotel neighbor dummy	-0.0123	-0.0107	-0.0143
		(0.0139)	(0.0133)	(0.0118)
	Has independent hotel neighbor dummy	-0.0041	-0.0019	-0.0061
		(0.0097)	(0.0098)	(0.0096)
	Has multi-unit owner hotel neighbor dummy	-0.0026	-0.0032	-0.0061
		(0.0113)	(0.0104)	(0.0095)
	Has a hotel neighbor managed by a multi-unit management company dummy	-0.0080	–	–
		(0.0136)		
	Has a hotel neighbor with no-affiliation dummy	–	-0.0124	–
			(0.0118)	
γ_j	City-level fixed effects?	YES	YES	YES
	Num. of observations	2931	2931	2931
	R-squared	0.11	0.11	0.13

*** p<0.01, ** p<0.05, * p<0.10
Heteroskedasticity robust standard errors in parentheses. All neighbor effects calculated for 0.5km radius.

Table 7: Specifications with 1-star reviews as dependent variable

		Difference in share of 1 stars reviews	Difference in share of 1 stars reviews	Difference in share of 1 stars reviews
X_{ij}	Site rating	-0.0087 **	-0.0092 **	-0.0100 **
		(0.0054)	(0.0054)	(0.0054)
	Hotel quality-tier controls?	YES	YES	YES
	Hotel age	0.00049 ***	0.0005 ***	0.0005 ***
		(0.00013)	(0.0001)	(0.0001)
Own_{ij}	Hotel is Independent dummy	0.011	0.012	0.0031
		(0.0096)	(0.0096)	(0.0123)
	Hotel is part of a large-owner chain dummy	-0.00078	-0.0023	0.0002
		(0.0048)	(0.0125)	(0.0049)
	Hotel is managed by a large management company dummy	-0.0061	–	–
		(0.0067)		
$NeighOwn_{ij}$	Has a hotel neighbor dummy	0.0152	0.0125	0.0101
		(0.0124)	(0.011)	(0.0104)
	Has independent hotel neighbor dummy	0.0191 **	0.0204 ***	0.0182 **
		(0.0078)	(0.0079)	(0.0079)
	Has large-owner chain hotel neighbor dummy	-0.0132	-0.0169 **	-0.0206 ***
		(0.0083)	(0.0079)	(0.0072)
	Has a hotel neighbor managed by a large management company dummy	-0.0140	–	–
		(0.0108)		
	Has a hotel neighbor with no-affiliation dummy	–	-0.0089	–
			(0.0087)	
γ_j	City-level fixed effects?	YES	YES	YES
	Num. of observations	2931	2931	2931
	R-squared	0.09	0.09	0.09

*** p<0.01, ** p<0.05, * p<0.10
Heteroskedasticity robust standard errors in parentheses. All neighbor effects calculated for 0.5km radius.

Table 8: Summary Statistics for Orbitz Reviews

	No. of Hotels	Mean	Std. Dev	Minimum	Maximum
Total Orbitz Verified Reviews	2569	32.59	47.75	1	628
Share of Verified Reviews 1 star	2569	0.093	0.16	0	1
Share of Verified Reviews 5 star	2569	0.241	0.20	0	1

difference between verified and unverified reviews are, nonetheless, interesting. For all Orbits verified reviews, 32% are 1s or 5s. For Orbitz unverified reviews, 58% are either 1s or 5s. Thus, unverified reviews are much more extreme than verified reviews. This, of course, could be due to unverified promotional reviews. However, it could also be the case that travelers who did not book through Orbitz nonetheless post reviews on Orbitz if they have had extreme experiences.

Table 9 repeats the regression specifications of Table 4, replacing Orbitz verified reviews with Expedia reviews. Regressions results are very similar to the results found in Table 4. As in Table 4, we find that Independent hotels have more 5 star reviews, and hotels from large ownership entities have fewer. In the Orbitz specification, the magnitude of the independence effect is somewhat larger than in our Expedia specifications, while the magnitude and significance of the multi-unit owner effect is smaller. Turning to 1 star reviews, we find, as in Table 4, that the presence of a neighbor has a positive but insignificant effect on 1 star reviews. Having an independent neighbor is associated with more 1 star reviews. As compared to Table 4, this effect is similar in magnitude, but is only statistically significant at the 15 percent confidence level. We also find, as in Table 4, that neighbors belonging to a large ownership entity are associated with fewer 1 star reviews. This effect is statistically significant at the ten percent level.

.

6 Conclusion and Limitations

We propose a novel methodology for empirically detecting review manipulation. In particular, we examine the difference in review distribution across Expedia and Tripadvisor, sites with different reviewer identity verification policies, and across different competitive/ownership conditions. Consistent with our theoretical claims, we find that an increase in hotel incentives to manipulate reviews results in an increase in our measures of manipulation. Substantively, we find that independent ho-

Table 9: TripAdvisor versus Orbitz Results

		Difference in Share of 5 star reviews	Difference in Share of 1 star reviews
X_{ij}	Site rating	-0.0076	-0.0029
		(0.0069)	(0.0048)
	Hotel age	-0.00095***	0.00044***
		(0.0002)	(0.00014)
	Hotel tier controls?	YES	YES
Own_{ij}	Hotel is Independent	0.048***	0.007
		(0.011)	(0.0106)
	Multi-unit owner	-0.013	-0.0030
		(0.009)	(0.0050)
	Has a neighbor	-0.0085	0.0109
		(0.012)	(0.0107)
$NeighOwn_{ij}$	Has independent neighbor	0.0076	0.0122
		(0.0106)	(0.0090)
	Has multi-unit owner neighbor	0.0188*	-0.0131*
		(0.0103)	(0.0084)
γ_j	City level fixed effects?	YES	YES
	Num of observations	2569	2569
	R-squared	0.04	0.05

***p<0.1, **p<0.05, *p<0.10
Heteroskedasticity robust standard errors in parentheses. All neighbor effects calculated for 0.5k radius.

tels engage in more review manipulation (both positive and negative), while hotels with multi-unit owners as well as hotels that are managed by a multi-unit management companies engage in less review manipulation (in the former case we find the effect for positive and negative manipulation, while in the latter we find the effect only in the case of positive manipulation). One important strength of our proposed methodology compared to earlier attempts is that our method does not require us to identify any particular review as fake or real, an inherently noisy and difficult task. Instead, we confine ourselves to examining differences between distributions.

Finally, we find that while the amount of review manipulation is economically significant, it is still small relatively to the total amount of reviewing activity. Why don't hotels engage in more intense review manipulation, given the fact that the mechanical costs of faking a review are low? Aside from any ethical concerns that the hotels have in engaging in this activity, we hypothesize that engaging in this activity exposes firms to reputational risks. The fact that the over-all level of manipulation activity seems to be relatively low is consistent with the notion that risks are perceived as relatively high. This perhaps explains how an open platform like Tripadvisor, that does not verify reviewer identity, can survive in the market. The obvious advantage of an open platform is that it allows the site to draw customers form all other sites (as well as from offline) as opposed to only restricting the reviews to its own customers. The downside is the degrading effect of review manipulation on the informational value of the site. Our empirical results show that the hotels are essentially able to self-police so that while they engage in some manipulation, the amount is not big enough to overwhelm the informational value of the site.

There are a number of limitations of this work. Perhaps the biggest limitation is that we do not observe manipulation directly but must infer it. This issue is of course inherent in doing research in this area. In the paper we deal with this limitation by building a strong case that the effects that we examine are due to review manipulation and not due to other unobserved factors. The second important limitation is that our measure of review manipulation does not include any content analysis. That is, one could imagine that one way in which a hotel could increase the impact of a fake review is by making particularly strong claims in the text of the review. For example, to hurt a competitor, a competitor could claim to be a traveler who witnessed a bed bug infestation. This is an interesting issue for future work.

Another limitation of this work is that we are unable to measure the impact that this manipulation has on consumer purchase behavior. Do consumers somehow detect and discount fake

reviews? Do they discount all reviews to some extent? Do they make poor choices on the basis of fake reviews? These questions are also left for future work.

7 Appendix

7.1 Proofs

Proof of Proposition 1:

The formulas in the Proposition are derived by taking the F.O.C.s of Equation (9) with respect to $e_{A,A}$ and $e_{A,B}$, and taking the F.O.C.s of (10) with respect to $e_{B,B}$ and $e_{B,A}$.

Endogenizing the Prices. As we argue in the main body of the paper, the firm does not expect manipulation to change its market share in expectation, given the optimal discounting by the consumer. Hence, the maximization problem in the second stage is the following:

$$\Pi^*_{A,Stage\,2} = \underset{p_A}{max} p_A \left[\frac{1}{2} + \frac{p_B - p_A}{2t}\right] - \delta_A \frac{\left(e^*_{A,A}\right)^2}{2} - \gamma_A \frac{\left(e^*_{A,B}\right)^2}{2} \tag{14}$$

$$\Pi^*_{B,Stage\,2} = \underset{p_B}{max} p_B \left[\frac{1}{2} + \frac{p_A - p_B}{2t}\right] - \delta_A \frac{\left(e^*_{B,B}\right)^2}{2} - \gamma_A \frac{\left(e^*_{B,A}\right)^2}{2} \tag{15}$$

After the appropriate substitutions (Proposition 1 provides $e^*_{A,A}$, etc.), taking the first order conditions, and some algebra, we have the following expressions for the equilibrium prices:

$$p_A = \frac{12t^3\delta_A\gamma_A\delta_B\gamma_B + 2t^2\delta_A\gamma_A(\delta_B+\gamma_B)\mu_s^2}{12t^2\delta_A\gamma_A\delta_B\gamma_B + 4\mu_s^2 t\left[(\gamma_A+\delta_A)\delta_B\gamma_B + (\gamma_B+\delta_B)\delta_A\gamma_A\right] + \mu_s^4\left[(\gamma_A+\delta_A)(\delta_B+\gamma_B)\right]} \tag{16}$$

$$p_B = \frac{12t^3\delta_A\gamma_A\delta_B\gamma_B + 2t^2\delta_B\gamma_B(\delta_A+\gamma_A)\mu_s^2}{12t^2\delta_A\gamma_A\delta_B\gamma_B + 4\mu_s^2 t\left[(\gamma_A+\delta_A)\delta_B\gamma_B + (\gamma_B+\delta_B)\delta_A\gamma_A\right] + \mu_s^4\left[(\gamma_A+\delta_A)(\delta_B+\gamma_B)\right]} \tag{17}$$

For simplicity, let's assume that $\delta_A = \gamma_A = \rho$ and $\delta_B = \gamma_B = 1$. We want to show that an increase in ρ (an increase in the reputational costs) results in less promotion on the part of firm A. Once

again, from Proposition 1 we know that in stage 3:

$$e^*_{A,A} = \frac{p_A \mu_s}{2\delta_A t}; e^*_{A,B} = \frac{p_A \mu_s}{2\gamma_A t} \tag{18}$$

$$e^*_{B,B} = \frac{p_B \mu_s}{2\delta_B t}; e^*_{B,A} = \frac{p_B \mu_s}{2\gamma_B t} \tag{19}$$

We take a derivative of these expressions, taking into account the fact that the prices are endogenous. That is, we can show that $\frac{\partial e^*_{A,A}}{\partial \rho} = \frac{\partial e^*_{A,B}}{\partial \rho} = \frac{\mu_s}{2t} \frac{\left[\frac{\partial p_A}{\partial \rho}\rho - p_A\right]}{\rho^2} < 0$.

Proof of Proposition 2:

Consider the firms' maximization problem:

$$\Pi^*_{A,Stage\,3} = \max_{e_{A,A}, e_{A,B}} \left(p_A \left[\frac{1}{2} + \frac{\mu_s(e_{A,A} + e_{A,B} - \hat{e}^*_{A,A} - \hat{e}^*_{A,B} + c_A) + p_B - p_A}{2t} \right] - \delta_A \frac{e^2_{A,A}}{2} - \gamma_A \frac{e^2_{A,B}}{2} \right) \tag{20}$$

$$\Pi^*_{B,Stage\,3} = \max_{e_{B,B}, e_{B,A}} \left(p_B \left[\frac{1}{2} - \frac{\mu_s(e_{B,B} + e_{B,A} - \hat{e}^*_{B,B} - \hat{e}^*_{B,A} + c_B) + p_A - p_B}{2t} \right] - \delta_B \frac{e^2_{B,B}}{2} - \gamma_B \frac{e^2_{B,A}}{2} \right) \tag{21}$$

The only difference here is that the consumer's inference ($\hat{e}^*_{A,A}$, $\hat{e}^*_{A,B}$, etc) will be different since the consumers can not observe the firm's cost function. Taking the derivative with respect to the promotion levels, it is clear that the optimal promotion level does not depend on the consumer's inference. That is, as before,

$$e^*_{A,A} = \frac{p_A \mu_s}{2\delta_A t}; e^*_{A,B} = \frac{p_A \mu_s}{2\gamma_A t} \tag{22}$$

$$e^*_{B,B} = \frac{p_B \mu_s}{2\delta_B t}; e^*_{B,A} = \frac{p_B \mu_s}{2\gamma_B t} \tag{23}$$

The consumer's inference will be a weighted average of the two types' optimal promotion levels, as is given in the Proposition. The result on share distortions follows directly from the fact that the consumer over-discounts the reviews for high-cost firm and under-discounts the review for low-cost

firm.

References

B. Anand and R. Shachar. (Noisy) Communication. *Quantitative Marketing and Economics*, 5(3): 211–237, 2009.

M. Anderson and J. Magruder. Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database. *The Economic Journal*, doi: 10.1111/j.1468-0297.2012.02512.x, 2012.

S. Anderson and R. Renault. Advertising Content. *American Economic Review*, 96(1):93–113, 2006.

K. Bagwell and G. Ramey. Advertising and Corrdination. *Review of Economic Studies*, 61(1): 153–171, 1994.

J. Berger, A. Sorensen, and S. Rasmussen. Positive Effects of Negative Publicity: When Negative Reviews Increase Sales. *Marketing Science*, 29(5):815–827, 2010.

R. Blair and F. Lafontaine. *The Economics of Franchising*. Cambridge University Press, 2005.

A. Chakraborty and R. Harbaugh. Persuasion by Cheap Talk. *American Economic Review*, 100(5): 2361–2382, 2010.

J. Chevalier and D. Mayzlin. The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research*, 43:345–354, 2006.

P. Chintagunta, S. Gopinath, and S. Venkataraman. The Effects of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation Across Local Markets. *Marketing Science*, 29(5):944–957, 2010.

V. Crawford and J. Sobel. Strategic Information Transmission. *Econometrica*, 50(6):1431–1451, 1982.

C. Dellarocas. Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management Science*, 52(20):1577–1593, 2006.

S. Dellavigna and E. La Ferrara. Detecting Illegal Arms Trade. *American Economic Journal*, 2(4): 26–57, 2010.

M. Duggan and S. Levitt. Winning Isn't Everything: Corruption in Sumo Wrestling. *American Economic Review*, 92(5):1594–1605, 2002.

Amy Harmon. Amazon glitch unmasks war of reviewers. *New York Times*, February 14 2004.

I. Horstmann and S. Moorthy. Advertising Spending and Quality for Services: The Role of Capacity. *Quantitative Marketing and Economics*, 1(3):337–365, 2003.

N. Hu, I. Bose, N. Koh, and L. Liu. Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Systems*, 52:674–684, 2012.

B. Jacob and S. Levitt. Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *Quarterly Jounral of Economics*, 118(3):843–877, 2003.

N. Jindal and B. Liu. Review Spam Detection. In *WWW '07 Proceedings of the 16th international conference on World Wide Web*, pages 1189 – 1190, 2007.

E. Kamenica and M. Gentzkow. Bayesian Persuasion. *American Economic Review*, 101(October): 2590–2615, 2011.

R. Kihlstrom and M. Riordan. Advertising as a Signal. *Journal of Political Economy*, 92(3):427–450, 1984.

L. Kornish. Are User Reviews Systematically Manipulated? Evidence from the Helpfulness Ratings. Leeds School of Business Working Paper, 2009.

D. Mayzlin. Promotional chat on the internet. *Marketing Science*, 25(2):155–163, 2006.

D. Mayzlin and J. Shin. Uninformative Advertising as an Invitation to Search. *Marketing Science*, 30(4):666–685, 2011.

M. Melnik and J. Alm. Does a seller's ecommerce reputation matter? Evidence from ebay auctions. *Journal of Industrial Economics*, 50(3):337–349, 2002.

P. Milgrom and J. Roberts. Price and Advertising Signals of Product Quality. *Journal of Political Economy*, 94(August):796–821, 1986.

B. Liu Mukherjee, A. and N. Glance. Spotting fake review groups in consumer reviews. *International World Wide Web Conference Committee*, April 16-20, 2012., 2012.

P. Nelson. Advertising as Information. *Journal of Political Economy*, 78(3):311–329, 1974.

M. O'Fallon and D. Rutherford. *Hotel Management and Operations*. John Wiley and Sons, 2010.

M. Ott, Y. Choi, C. Cardie, and J. Hancock. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 309–319, 2011.

P. Resnick and R. Zeckhauser. *Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System*, volume 11, pages 127–157. 2002.

P. Resnick, R. Zeckhauser, J. Swanson, and K. Lockwood. The value of reputation on eBay: A controlled experiment. *Experimental Economics*, 9(2):79–101, 2006.

Scott and Orlikowski. Reconfiguring relations of accountability: Materialization of social media in the travel sector. *Accounting, Organizations, and Society*, 37:26–40, 2012.

I. Vermeulen and D. Seegers. "tried and tested: The impact of online hotel reviews on consumer consideration". *Tourism Management*, 30(1):123–127., 2009.

R. Law Ye, Q. and B. Gu. The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, 28(1):180–182, 2009.

N. Zmuda. Ann taylor investigation shows ftc keeping close eye on blogging. *Advertising Age*, April 2010.

Grab them Before they Go Generic: Habit Formation and the Emerging Middle Class*

Alon Eizenberg†
Department of Economics
Hebrew University of Jerusalem

Alberto Salvo‡
Kellogg School of Management
Northwestern University

July 7, 2012

Abstract

The "emerging middle class" has become a force of great economic importance in consumer markets around the globe. This paper examines the impact of a substantial rise in Brazil's living standards on the development of the country's large soft-drink market, during a six-year period which saw unprecedented growth in the share of generic soda brands. Combining richly varying market and consumer-level data, we estimate a novel structural demand model that identifies a mechanism by which a household develops either a "premium brand habit" or a "frugal habit." We find strong empirical evidence of such persistence in preferences. Our results demonstrate that habit formation plays a crucial role in this emerging market: the arrival of many new consumers, who have not yet developed established habits, allows generic producers to more easily tap into this new demand for soda. Moreover, this persistence in preferences provides strong justification for Coca-Cola Co's decision to abruptly cut prices in July 1999. An estimated variant of our model, that does not account for habit formation, provides much weaker support for this strategic price cut.

Keywords: emerging middle class, social mobility, differentiated-product demand, habit formation, generics, competitive fringe, premium brands

JEL Classification: L10, D12, O12

*A previous draft was circulated under the title "Demand in the Wake of an Emerging Middle Class and Low-End Entry." For facilitating access to data, we wish to thank Ricardo Fort, Bruno Gouvea, Isadora Nardy, Claudia Pessoa and Daniela Pisetta. We also thank Alaor Dall'Antonia Junior and Maria Cristina Costa at the National Institute of Meteorology (INMET) for access to their data. We thank Maria Ana Vitorino for discussing the marketing literature with us. We are grateful to Itai Ater, David Besanko, Allan Collard-Wexler, Pierre Dubois, David Genesove, Noam Gruber, Paul Grieco, Saul Lach, Katja Seim, and Scott Stern, as well as seminar/conference participants at the Economic Workshop at IDC Herzlia, the I.O. day at Tel Aviv, the IIOC, the Northwestern-Toulouse workshop, the annual meeting of the Israeli Economic Association, Hebrew University of Jerusalem, Kellogg Marketing (KAMP), Oxford University, Tel Aviv University, and the University of Virginia for valuable comments. All errors are our own.

†aloneiz@mscc.huji.ac.il

‡a-salvo@kellogg.northwestern.edu

"A study this year by the United Nations Economic Commission for Latin America and the Caribbean concluded that tens of millions of the region's inhabitants have risen into the middle class over the past two decades. That's prompted 'a notable expansion of the consumer market,' ...(thanks to) the prospects of los emergentes—the emerging ones—as marketers call the newly minted middle-class members."

Matt Moffett, *The Wall Street Journal*, November 15, 2011

"Across the developing world millions—perhaps billions—of people are currently forming tastes that will endure for the rest of their lives. Put one of Kraft's Oreos or Cadbury's Flakes in their hands and they may become loyal customers for decades to come."

The Economist, November 5, 2009

1 Introduction

The "emerging middle class" has become a major economic phenomenon in consumer markets around the globe. Since the mid 1990s, many developing countries, as far-flung and varied as Brazil, China, India, Indonesia and Turkey, are experiencing a socioeconomic transformation, whereby a substantial mass of low-income households emerge from below the poverty line and begin to consume goods and services that they previously could not afford.[1] Bolstering the demand for many consumer goods, these "new consumers" provide a potential engine of growth for the global economy. This new source of demand calls for the development of empirical tools to examine both its nature and the implications for competition in emerging consumer markets.

Our paper examines this demand expansion process via an important test case: the Brazilian market for carbonated soft drinks (or "soda"). We study the evolution of this market from December 1996 through March 2003, a six-year period over which two striking phenomena were evident: a substantial expansion in demand fueled by rising living standards, and the rapid growth of a competitive fringe of soda producers.

Brazil's large soda market trails only the United States and Mexico by volume. Following a successful economic stabilization plan in 1994, aggregate soda consumption doubled by 1997, and continued to grow at an annual rate of about 10% through 1999. As is well documented, this growth was driven by pronounced upward mobility among lower income households, who

[1] The Economist (2011a) states that, using a broad income definition, "(t)he middle classes...trebled in number between 1990 and 2005 in developing Asia to 1.5 billion." Nomura Bank states that by 2014 Indonesia should boast almost 150m "newly affluent Indonesians (who) are certainly spending" (The Economist, 2011b).

were no longer forced to pay an "inflation tax."[2] In 1999, the Financial Times reported that "the increased purchasing power that came with stable prices... allowed about 25m new consumers into the (soft drink) market" (among a total population of 170m at the time). Other markets, ranging from fresh meat to refrigerators to housing, saw similar expansions in demand.

One might have expected that established soda producers, namely the Coca-Cola Company (hereafter Coca-Cola) and Ambev, who in 1996 jointly accounted for almost 90% of Brazilian soda expenditure, were best positioned to tap into this new demand for soda.[3] Instead, between 1996 and 1999, the combined volume share of (ultimately) hundreds of regionally-focused discount brands—which we label "generics"—doubled from 20% to 40%.[4] In contrast to the dominant duopoly's heavy investments in advertising, generic producers focused their marketing efforts on securing shelf space via low prices. With the stiff competition slowing down company growth, "Coca-Cola blamed difficulties in developing countries such as Brazil when it shocked Wall Street in December (1998) by announcing a rare drop in quarterly sales" (Financial Times 1999).

Having kept prices broadly constant during the preceding years of entry and expansion in the fringe, in 1999 Coca-Cola abruptly cut prices across its brands by over 20%, a move that was soon matched by Ambev. Following this price cut, the growth in the market share of the generic fringe was halted. As we discuss below, however, the fringe was able to hold its ground, continuing to command substantial market share even after the premium brands' large price cut.

The goal of this paper is to examine whether—and via which mechanisms— the emerging middle class can provide fertile ground for the growth of a generic fringe.[5] We focus on two possible (and not mutually exclusive) mechanisms that may have been playing a role in the Brazilian soda market. First, emerging middle-class consumers may have been price sensitive, and thus likely to favor cheap generics over expensive brands. To stay with the Financial Times' analysis, "(t)he new (soda) customers...had different priorities...(t)hey were less concerned about expensive TV ads and more interested in value." A price-sensitive, expanding consumer segment may help explain both the growth of generics and the premium sellers' price cut.

Second, it is conceivable that, upon their arrival in the market, the new customers were starting to form consumption habits and tastes in the soda category. The absence of habits may have aided generic entrants in making inroads into this emerging consumer segment. Moreover, if

[2] A substantial mass of households with no access to inflation-indexed bank accounts were the main beneficiaries of the taming of chronically high inflation: "...Jose Benevenuto, a 53-year-old Rio de Janeiro bus driver...still recalls the years in the early 1990s when Brazil's four-digit inflation forced him to rush to the supermarket as soon as he was paid so he could spend his money before it lost all value" (Wall Street Journal 2011). By 1995, inflation was (sustainably) down to single-digit annual levels.

[3] Ambev distributed the Pepsi brand, and is now part of the AB Inbev group.

[4] This pertains to the dominant market segment of family-size bottles sold through the "self-service outlets" distribution channel (supermarkets with checkouts) in urban areas.

[5] Several emerging markets appear to feature a substantial presence of generic producers, underscoring the research question. The China-based appliance manufacturer Galanz cites the National Bureau of Statistics in claiming that there were "nearly 300 brands in (the) Chinese market" in 2008 (Galanz 2008). The Economist (2012) counts 100 "domestic carmakers" in China. Abbott India's brands Digene, Eptoin and Cremaffin face competition from 211, 327 and 242 "regional" generics, respectively (as shared by the company during a corporate presentation in late 2011).

persistence in consumer tastes is important, this second mechanism may have provided a strong incentive for Coca-Cola to cut prices with the goal of defending its future market position.[6]

The extent to which the emerging middle class is price sensitive, as well as the extent to which habit formation plays a role in such markets, are matters for empirical investigation. To this end, we develop and estimate a structural model of demand that allows us to segment consumers according to both their socioeconomic standing and their consumption habits. Our model is well-suited for a fast-changing emerging market setup, and is estimated using a combination of market-level and consumer-level data. Building on the random-coefficient logit framework, our model displays two novel features.

First, we allow consumers to belong in one of three discrete demographic groups: "poor," "established affluent," or "*newly* affluent." Established affluent households are those who were already affluent before the process of upward mobility began, whereas newly affluent households represent the new middle class. In our model, poor households can move up to newly affluent status, while downward mobility is captured by allowing newly affluents to move down to poor status. Such downward mobility is apparent toward the later part of our sample period, when the Brazilian economy was hit by a recession. In addition to upward and downward mobility, our model also accounts for *urbanization*, another pervasive demographic shift.

The second key component of our model is habit formation, of a special kind. In particular, we allow for three habit states: a habit to consume premium soda brands (a "premium habit"), a habit to consume generics (a "generic" or "frugal" habit), or not developing a habit to consume soda. Habits develop according to the choice made by the household in the immediately preceding period.[7] We label this a *Brand Type Persistence* (BTP) mechanism, as it captures a persistence in demand for a certain type of good (i.e., premium or generic). In our model, developing a premium habit in period $t-1$ (by consuming, say, Coke) increases the utility from consuming any premium brand (say, Coke, Fanta or Pepsi) in period t. Similarly, recent consumption of a generic brand raises the utility from current consumption of this or any generic brand. This parsimonious modeling approach allows us to capture the key dichotomy between premium versus generic soda products in a rapidly evolving market, where consumers establish shopping patterns that may endure into the future.

A household's type in our model is determined by both its current socioeconomic standing and its current habit state. We extend the empirical literature on discrete-type demand models to our emerging market setup. In particular, we develop an estimation algorithm that tracks the

[6]The role of habit formation in food and beverage has been emphasized in the literature. See Atkin (2011) and Bronnenberg, Dubé and Gentzkow (2012) for recent contributions.

[7]Much of the extant literature on state-dependent preferences assumes that current "habits" were developed in the immediately preceding period. In most papers, periods are captured as shopping trips in household-level retail scanner data. In contrast, a period in our setting lasts one or two months, an interval we view as more appropriate for our habit formation context. A more general model would allow habits to evolve as a function of consumption choices over multiple preceding periods.

population fractions that belong in each type over time. This is accomplished by combining data on aggregate demographic trends with brand choices predicted from our utility framework.

Findings. We find that, while newly affluent households are significantly less price sensitive than the poor, their price sensitivity is comparable to that of the established affluents. Our findings do not, therefore, lend support to the notion that the new middle class was especially price sensitive, and that it was via this mechanism that the emerging middle class provided suitable conditions for the growth of the competitive fringe.

Our estimated model does, in contrast, provide strong empirical evidence for the second mechanism hypothesized above: habit formation. This persistence in preferences is shown to be of both statistical and economic significance. Our model allows us to evaluate the *monetary value of habit formation*: for example, a "frugal habit" increases a newly affluent consumer's willingness to pay for a generic product by (Brazilian Real) R$ 2.04 (about US$ 1) per liter relative to displaying no habit. This suggests that habit formation played a prominent role in driving the growth of generic brands. It also explains the sense of urgency with which premium brands acted when they cut prices in mid 1999: had they failed to cut prices, an increasing fraction of new middle-class consumers would have "gone generic." A premium price cut helped ensure that many of these consumers developed a premium habit instead.

We further demonstrate this intuition by employing our model in counterfactual analysis. We find that, had premium brands failed to cut prices in mid 1999, they would have seen their market shares and variable profits suffer substantial declines through 2003. Our estimated BTP model, therefore, provides strong justification for this strategic move. This analysis further indicates that the price cut was more effective with consumers who were yet to form soda-consuming habits, and less effective with consumers who had already developed a generic habit. Finally, and importantly, an estimated model variant that shuts down the habit mechanism provides much weaker support for the premium price cut.

Identification. Our empirical approach faces a familiar challenge: how can one separately identify consumer heterogeneity from persistence in tastes? In our analysis, this is accomplished by relying on the rich cross-sectional and time-series data variation in this rapidly changing market. In particular, we exploit region-specific social mobility and pricing variation that is likely exogenous to demand *unobservables*. For example, the magnitude and abruptness of Coca-Cola's nationwide price cut halfway into the sample period strongly suggests that it was unlikely to be correlated with any contemporaneous region-specific shocks to demand, making the price cut itself an effective instrument (Salvo 2009).

Our model allows price sensitivity to vary at the socioeconomic level, identifying it off of the observed co-variation of socioeconomic shifts (i.e., upward and downward mobility), prices and market shares. Identification of the habit mechanism also follows from data variation: for

instance, during the recession that set in toward the end of our sample period, households fell back from newly affluent status to the ranks of the poor, yet soda consumption did not fall.

Our framework offers another important insight regarding identification: failing to control for persistence may frustrate the identification of the distribution of consumer price sensitivity. Indeed, the model variant that we estimate shutting down habits biases the price sensitivity of the newly affluent in the direction of that of the poor. Intuitively, this may be explained by the recessionary period, when an increasingly poor population continued to consume stable amounts of soda. A model that does not allow for persistence would have to interpret this as evidence that the poor are "not that price sensitive."

Literature. Our study contributes to different lines of research. One line examines competition between branded products and lower cost generics, particularly in pharmaceuticals, including Chaudhuri, Goldberg and Jia (2006) in India, and Hurwitz and Caves (1988) and Scott Morton (2000) in the US. Another line of work examines the relationship between the demographic composition of demand and prices, or inflation moderation (e.g., Frankel and Gould 2001, Bils and Klenow 2004, Nevo and Hatzitaskos 2006, Lach 2007, Calzolari, Ichino and Manaresi 2012).

The empirical literature in economics and marketing has introduced habits or persistence into models of consumer choice, including Eichenbaum, Hansen and Singleton (1988), Erdem (1996), Keane (1997), Shum (2004) and Dubé, Hitsch, Rossi and Vitorino (2008). Our work differs from the extant literature along several dimensions. First, existing studies tend to rely on micro-level panel data, repeatedly observing an individual household's purchasing behavior. When studying an emerging market, repeated observations on a fixed panel of households are less likely to be available. Furthermore, they may miss the demographic shift that lies at the heart of the analysis. Our paper demonstrates that a model with persistent preferences can be estimated with a panel of market-level data (in addition to a single cross-section of household-level data). This is made possible by the rich data variation afforded by the emerging market setup.

Second, the state dependence we model differs from brand loyalty. It is motivated by a desire to address the heterogeneity in business models between premium and generic sellers which plays an important role in some emerging markets.[8] By emphasizing this aspect, our goal is to capture a potentially important mechanism in such settings, rather than to extend the brand loyalty literature. Our focus on emerging markets rather than mature ones marks another departure from the extant literature on persistent preferences. Beyond its economic importance, the emerging market setting provides a unique opportunity to identify and study state-dependent preferences. Finally, this paper wishes to contribute to a better understanding

[8]Executives of a global "fast-moving consumer goods" firm meeting one of us recently in Delhi stated that "as a company in the *A business* we don't naturally understand the *B business*, where the *value* proposition is at the heart of it, putting us at a certain disadvantage when selling to the *Bottom of the Pyramid* in the Indian market." (To be clear, all words—including the terms in italics—are the executives' own, though in slightly rearranged order without modifying context.)

of demand in emerging market settings. Another example is Sancheti and Sudhir (2009), who examine the consumption of education in India. The rapid growth of such markets suggests that studying them offers a promising avenue for applied microeconomic research.

The rest of the paper is organized as follows. Section 2 describes the data and the joint phenomena that motivate our analysis: upward social mobility and the growth of the generic fringe. Section 3 develops our demand model. Section 4 explains our estimation algorithm and provides arguments for identification. Section 5 reports estimation results, as well as counterfactual analysis. Section 6 concludes.

2 Market and data

This study brings together data from three main sources. The following subsections describe these data sources, as well as the manner with which they reflect the two striking phenomena discussed above: the emergence of a new middle class and the growth of the generic fringe.

2.1 Market-level data

We observe a panel of market-level data from Nielsen, consisting of total quantities and prices for soft drink brands. There are $g = 1, ..., 7$ regions and $t = 1, ..., 57$ time periods, ranging from the December 1996-January 1997 bimonth to the March 2003 month (Nielsen raised the frequency of its bimonthly point-of-sale audits to a monthly basis in 2000). We therefore observe $7 \cdot 57 = 399$ region-period markets.

The seven geographic markets are urban and, as in Salvo (2009), we consider soft drinks sold through the "self-service" channel (supermarkets with checkouts) in the 2-liter family-size bottle. Our focus on this market segment is justified on several counts. First, the focus on urban areas is natural since more than 80% of Brazil's population was urbanized by 1996 and, importantly, our framework allows for rural-to-urban migration. Second, urban households in Brazil perform most of their grocery shopping in supermarkets with checkouts, rather than in traditional behind-the-counter retail stores. Finally, sales of family-size bottles dominate those of "single-serve" (300ml) bottles or cans (mostly sold in bars and restaurants). Moreover, the competitive fringe, whose success we wish to explain, was mostly present in the family-size bottle segment.

Also following Salvo (2009), we aggregate flavors and brands into $j = 1, ..., 9$ brand-groups. These groups include eight "premium brands" (or "A-brands"): five brands of the Coca-Cola Company (Coke, Fanta, the guaraná-flavored Kuat, Diet Coke, and "Other Coca-Cola"), and three brands marketed by Ambev (Guaraná Antarctica, Pepsi, and "Other Ambev"). The ninth brand category is an aggregate of discount brands (or "B-brands") that form the generic fringe.[9]

[9]The data provide limited information on the breakdown of this group into individual discount brands, as they are so numerous.

Table 1 describes the volume shares (of the soda category) for each of the nine brands across the seven Nielsen regions, in the first and last periods in our sample (all statistics pertain to family-size bottles sold in supermarkets with checkouts). Averaged arithmetically across regions, Coca-Cola's brands accounted for a 50% volume share in the first period, with Coke being dominant, whereas Ambev enjoyed a 31% share, with Guaraná Antarctica and Pepsi as its flagship brands. The table reports the stark growth in the generic share, from 19% at the start of the sample to 40% at the end. The table reflects some region-specific tendencies to consume particular brands. Our empirical framework controls for such region-brand effects.

Defining market size. We denote the observed quantity and price associated with brand j sold in the region-period market gt by q_{jgt} and p_{jgt}, respectively. As is common in discrete-choice applications, we need to define the size of market gt, that is, the maximum amount of soft drinks that can potentially be consumed in this market. We define this quantity, denoted $\mathcal{M}_{gt}$, as six liters per week over the duration of period t multiplied by the number of urban households residing in market gt (which we obtain from a fourth data source). One may interpret the six liters per week as three weekly family meals in which a 2-liter family-size bottle of soda might be brought to the table (rather than water, juice, etc). We then compute brand j's share as $s_{jgt} = q_{jgt}/\mathcal{M}_{gt}$. The share of the outside option (that is, the option not to consume soft drinks) is given by $s_{0gt} = 1 - \sum_j s_{jgt}$.[10]

The growth of the competitive fringe and Coca-Cola's response. In contrast to the established Coca-Cola/Ambev duopoly, with their heavily advertised brands and nationwide distribution, fringe players ran small-scale operations, in most cases individually covering a fraction of a state, and selling at substantially lower prices. Having hovered around a 15% volume share of the soda category at least since 1980 (Salvo 2009), the fringe began growing strongly in the mid 1990s, as evidenced in Table 1. A shift from the returnable proprietary glass bottle (returned to the bottler for reuse, requiring a certain level of sophistication and scale) to the inexpensive non-returnable 2-liter PET bottle may have lowered barriers to entry (Ambev 2003). No census of fringe operators exists, but industry sources suggest that following three years of substantial entry, the number of firms selling generic soda may have reached 500 by 1999.[11]

Figure 1 reports that both premium and generic brands enjoyed substantial volume growth during the sample period (for illustrative purposes only, the figure aggregates quantities sold over all seven regions, and aggregates all eight premium brands together). Importantly, the generic fringe grew much faster than the premium brands over the first 30 months of the sample—that is, until Coca-Cola's abrupt mid 1999 price cut. The figure also reveals strong seasonality effects,

The aggregate structure of this data is indicative of how Coca-Cola and Ambev—two of Nielsen's largest customers—viewed brand differentiation within the generic fringe.

[10] The appendix reports robustness checks which reassure us that our results are not overly sensitive to the market size definition.

[11] The Financial Times (1999) speaks of over 900 companies in the industry, while Ambev's (2003) financial reports refer to 700 "low-price brands."

for which we control in the empirical application.

The left panel of Figure 2 illustrates the evolution of (mean share-weighted) prices for premium brands and for generics in R$ per liter.[12] Premium brands initially held prices broadly flat, at R$ 1.15. In mid 1999, Coca-Cola cut its prices by more than 20%, a move that was soon matched by Ambev. The figure clearly indicates the sudden nature of this price cut, which we will exploit for identification purposes.

For their part, prices in the fringe declined gradually but relentlessly, from R$ 0.90 in late 1996 to R$ 0.60 in late 2000.[13] Falling generic prices are consistent with substantial entry and capacity expansion in the fringe, as competitive firms passed efficiency gains through to consumers. Fringe prices did not respond to the premium price cut in the sense that they did not deviate from their trend, consistent with competitive behavior.

As the right panel of Figure 2 shows, after 30 months of generics gaining share at the expense of the premium brands, the premium price cut had a clear and immediate impact. It essentially put an end to the staggering generic growth, and led to stable volume shares for premium and generic brands through the end of the sample.

2.2 Data on aggregate social mobility

To track the undercurrent of social mobility in the Brazilian economy, we rely on the proprietary LatinPanel survey from IBOPE, a leading (private-sector) provider of data on consumer demographics.[14] The survey, widely used by marketing practitioners, profiles urban households in Brazil's different regions based on their expenditure on durable goods and services (e.g., ownership of a refrigerator, numbers of TVs and bathrooms in a residence, current employment of house maids, education attainment). Adopting an industrywide points scale (ABEP 2003), each household is assigned to a "socioeconomic group." The IBOPE data that we have access to covers the period 1994-2006 (with 1995 missing) and provides the proportion of urban households that belong in either the **AB**, **C** or **DE** groups (respectively with "high," "intermediate" or "low" levels of affluence) in each of seven geographic regions.[15]

The IBOPE data indicate that the demographic composition of urban households: (i) was stable between 1994 and 1996; (ii) displayed strong upward mobility from **DE** to **ABC** (i.e., {AB,C}) status between 1996 and 2000; and (iii) experienced a partial reversal of this upward

[12] Throughout the paper, R$ prices are reported at constant Brazil CPI March 2003 terms (divide by 2 for rough US$ values).

[13] Given that we convert prices to constant R$ (see the appendix), what this means in practice is that nominal prices in the fringe fell 17% compared with the overall price level in the economy (the CPI) growing by 25% over the 45 months to September 2000, i.e., $.6/.9 \simeq (1 - .17) / (1 + .25)$.

[14] The company's name is so established among Brazilian households that, as cited in Wikipedia, it is synonymous with research (e.g., see the *Aurélio* Portuguese language dictionary). Coca-Cola kindly shared the data with us for the purpose of this study.

[15] The points scale used to classify each household stays clear of income, there being reasons why income-based measures might less accurately reflect changes in the standard of living (Carvalho Filho and Chamon 2011, Economist 2007). That said, to provide perspective, mean annual incomes in 2000 for **C** and **DE** urban households were respectively US$ 6,100 and US$ 2,600 (ABEP 2003, based on an IBOPE survey, using nominal 2000 R$/US$).

mobility thereafter, consistent with a recession setting in at that time. In aggregate, the proportion of **DE** households fell from 50% in 1996 to 33% in 2000, then rose to 44% by 2003 (conversely, the **AB** proportion rose from 19% in 1996 to 33% in 2000, then fell to 23% by 2003).

These demographic patterns are consistent with media and market research reports. The 1996-1999 upward mobility was fueled by successful economic reforms in the early 1990s, including trade liberalization and, most notably, the taming of very high inflation by the 1994 *Real* stabilization plan. These reforms were followed by strong consumption growth across the Brazilian economy, particularly among lower-income households. Figure 3 reports per capita consumption between the mid 1980s and mid 2000s in two different sectors—beverages (soft drinks) and housing (cement); a similar temporal pattern leading up to 2000 is present.[16]

The Boston Consulting Group (2002), reporting on its own household survey, spoke of the emergence of a middle class with "very strong consumer potential," whereas Fátima Merlin, chief economist for the Brazilian Association of Supermarkets (ABRAS), referencing the same IBOPE data that we use, stated that "following the *Real* Plan, thanks to price stability and real growth in workers' earnings, consumer markets experienced entry by households previously outside such markets, *with upward migration from the 'E' and 'D' segments of the population* to the 'C' segment, *as the IBOPE data indicate*" (SuperHiper 2003; emphasis added).

As for the downward mobility reported by IBOPE over 2001-2003, economic episodes that may have dampened investor and consumer sentiment include the 1997-98 Asian crisis, the 1999 Brazilian currency crisis, and the 2000-01 Argentine crisis.[17]

To analyze the impact of the changing socioeconomic composition, we define three socioeconomic groups: "Established Affluent" (EA), "Newly Affluent" (NA) and "Poor" (P). Using the IBOPE proportions together with urban household counts, we track the number of households who belong in each of these groups, in each region and over time.[18] Our "Established Affluent" group consists of urban households who were already in **ABC** status in 1996, i.e., before the process of upward mobility took off. The number of households in this group, in each of the seven regions, is thus fixed across time at the initial number of **ABC** households in that region. We define the size of the "Poor" group in each region-period market gt by that market's number of urban households who belong to socioeconomic group **DE**.

Finally, we define the size of the "Newly Affluent" group in market gt as the difference between the contemporaneous number of **ABC** households and region g's initial (i.e., 1996) number of **ABC** households. In other words, the number of time-t newly affluents is computed by subtract-

[16] See Carvalho Filho and Chamon (2011) and Salvo (2009, 2010) for further discussion of the consumption effects of reforms in the 1990s. See also Neri (1995).

[17] A similar temporal pattern of prosperity can be detected in earnings data in IBGE's monthly survey of earnings and employment, conducted in 6 large cities, though the turning points in the series tend to occur sooner than 2000. Details are available from the authors upon request.

[18] The appendix details how we interact IBOPE's urban socioeconomic distributions with the number of urban households from IBGE's annual household surveys (PNAD), as well as consistency checks between IBOPE and IBGE survey data.

ing the region's (fixed) number of established affluents from the number of time-t households in **ABC** status.[19]

To illustrate our computations by way of an example from the IBOPE data, in the South region there were: (i) in $t = 1$ (Dec-96/Jan-97), 3149 (thousand urban) **ABC** households and 2116 **DE** households, and (ii) in $t = 2$ (Feb/Mar-97), 3238 **ABC** households and 2045 **DE** households. Between these periods, $3238 - 3149 = 89$ poor households moved up to newly affluent status (and the number of migrants grew by $3238 + 2045 - (3149 + 2116) = 18$). Thus the numbers of established affluents, newly affluents and poor in this region, respectively, are (3149,0,2116) in period 1 and (3149,89,2045) in period 2.

In the data, the number of newly affluent households is strictly positive for all regions and all time periods $t > 1$, and is equal to zero, by definition, for $t = 1$, the initial period of our Nielsen soda market data described above. The zero number of newly affluents in period 1 is justified by the fact that, in the IBOPE data, the process of upward mobility takes off just before our Nielsen sample begins in late 1996. This assumption is also consistent with press and trade articles from the time. For example, our measure of the number of newly affluent households in 1999, summed across the seven Nielsen regions, translates into 20m consumers, a notch below the Financial Times' (June 1999) count of "(Brazil's) 25m new consumers (in the aftermath of an) economic plan" (and noting that our study does not cover rural areas or the northern states).

Figure 4 plots the evolution of the socioeconomic composition by region, i.e., the population fractions of established affluent, newly affluent and poor households. The figure clearly demonstrates the emergence of a new middle class. The increase, toward the end of the sample period, in the fraction of the poor at the expense of the fraction of newly affluents reflects the joint effects of the recession and the urbanization process. There are large regional disparities, with region 1 (states in the Northeast) being the least affluent and region 4 (São Paulo Metro) being the most affluent (65% and 36% of urban households in these regions are initially poor, respectively).

2.3 Data on household-level brand choices

Our third main data source allows us to relate household characteristics to soda consumption choices at the beginning of our period of study. We use an urban household expenditure survey conducted between October 1995 and September 1996 by IBGE (a federal agency equivalent to the US Census Bureau and Bureau of Labor Statistics combined). This survey (hereafter HEX 95/96) reports the type of soda brand purchased, as well as the amount spent, for consumption inside the home. Households in the survey are not classified according to the ABCDE system, but we use the detailed information available (e.g., ownership of a refrigerator, numbers of TVs

[19] In addition to upward and downward mobility, another demographic force affecting group sizes is (net) rural-to-urban migration. We capture this process via changes in the urban household population, assuming that households migrating to the city are initially "Poor." See Assumption 2 below.

and bathrooms in the residence, current employment of house maids, education attainment) to assign, like IBOPE does, each household to a socioeconomic group from A to E.

Table 2 reports the relationship between inside-the-home consumption of soft drinks and socioeconomic status. For example, 34.5% of São Paulo Metro's (region 4) **ABC** households in 1996 purchased soda for home consumption whereas only 19.8% of **DE** households did so. Table 2 also shows, for each of the different regions, the share of **ABC** households who consume a premium (generic) brand, and similar figures for **DE** households. These reflect the co-variation between a household's socioeconomic standing and its choice between premium and generic brands. Across all cities, **DE** soda-consuming households were more likely to purchase generics over premium brands (12 : 163) *relative* to **ABC** soda-consuming households (11 : 339). It is worth noting that our modeling of soda-consuming households at each point in time as either premium or generic shoppers, but not "hybrids," is largely consistent with the HEX data.

As we explain below, the fact that the HEX survey was conducted shortly before the beginning of our Nielsen market data allows us to use this information as an "initial condition" for the evolving relationship between socioeconomic standing and consumption choices.

Additional data sources. Our analysis draws on additional data: (i) the population of urban households by region from IBGE's expanded annual household surveys (PNAD), (ii) proprietary McCann-Erickson data on advertising intensity at the brand-market level, (iii) proprietary temperature data (another demand shifter) from the National Institute of Meteorology, and (iv) data from various sources on cost shifters such as the prices for sugar, electricity and fuel.

3 The model

We develop a model of household demand for soft drinks which accounts for socioeconomic standing and habit formation. Our model accommodates the fundamental features of the data noted above. In particular, we do not observe household-level data over time. We do observe the region-specific, temporal evolution of aggregate brand market shares and prices, and of households' socioeconomic composition. The proposed model allows us to identify consumer demand, and, importantly, the persistence component, given the available data variation.

3.1 Household types and the utility framework

In each period t, a household belongs in one of three socioeconomic groups (EA, NA, P) (again, established affluent, newly affluent or poor) and, consistent with the data, we allow for region-specific, aggregate mobility across these groups over time.

We denote the eight premium brands (or A-brands) as elements of the set $\mathcal{A}$, and the ninth brand category as the only element of the set of generics (or B-brands) $\mathcal{B}$. A household's current

preferences over substitute soft drink brands depend on its current socioeconomic standing as well as on the household's previous-period consumption by virtue of a habit mechanism. We allow for three habit states. Specifically, we differentiate households who, in the preceding period, consumed: (i) a premium brand $j \in \mathcal{A}$, (ii) a generic brand $j \in \mathcal{B}$, or (iii) did not consume soda at all, i.e., chose $j \in \mathcal{O}$, with $\mathcal{O}$ denoting the outside option set. Crossing together the three socioeconomic states and the three habit states, we obtain nine discrete household types, indexed by r:

$$r \in \mathcal{R} := \left\{ EA^{\mathcal{A}}, EA^{\mathcal{B}}, EA^{\mathcal{O}}, NA^{\mathcal{A}}, NA^{\mathcal{B}}, NA^{\mathcal{O}}, P^{\mathcal{A}}, P^{\mathcal{B}}, P^{\mathcal{O}} \right\} \tag{1}$$

Thus, for example, a time-t newly-affluent household who consumed a generic brand in period $t-1$ is of type $r = NA^{\mathcal{B}}$, whereas an established affluent household who consumed a premium brand in the preceding period is of type $r = EA^{\mathcal{A}}$. Fixing a region-period market gt, let $F_{r,gt}$ denote the fraction of that market's household population that belongs to type r. We collect these fractions for the nine types in a 9-dimensional vector denoted $\mathcal{F}_{gt}$, to which we refer as market gt's *type-distribution vector*.

The indirect utility of household i of type r in market gt from consuming brand j is given by:

$$u_{i \in r,j,gt} = \delta_{jgt} + \alpha_r \cdot p_{jgt} + \lambda \cdot h_{jr} + \epsilon_{ijgt} \tag{2}$$

We now explain each component of this function. The term δ_{jgt} denotes a market-specific, household-invariant base utility from brand j:

$$\delta_{jgt} = x'_{jgt}\beta + \alpha \cdot p_{jgt} + \xi_{jgt},$$

where x_{jgt} contains brand-region fixed effects, seasonal effects, brand-level advertising, market temperature, and region-specific time trends. These trends allow for region-specific temporal evolution in the utility from the outside option, such as differential rates of expansion in markets for soft-drink alternatives (e.g., juices). The brand's price is p_{jgt}, and ξ_{jgt} denotes a (brand-market specific) utility shock observed by firms and consumers, but unobserved to the econometrician, and (α, β) are coefficients to be estimated.

The second and third terms in (2) introduce household-type heterogeneity. The parameter α_r shifts the base price sensitivity α in accordance with the household type r:

$$\alpha_r := \begin{cases} \alpha_{EA} \text{ if } r \in \left\{EA^{\mathcal{A}}, EA^{\mathcal{B}}, EA^{\mathcal{O}}\right\} \\ \alpha_{NA} \text{ if } r \in \left\{NA^{\mathcal{A}}, NA^{\mathcal{B}}, NA^{\mathcal{O}}\right\} \\ 0 \text{ otherwise} \end{cases}$$

This implies that while α is the price sensitivity of poor households, the sums $(\alpha + \alpha_{EA})$ and $(\alpha+\alpha_{NA})$ are the price sensitivities of the established affluent and the newly affluent, respectively. Note that we allow price sensitivity to vary with a household's socioeconomic standing, but not with its "habit." Below we provide intuition for the role played by this restriction in the identification of the model.

The variable h_{jr} in (2) captures the persistence, or "habit," feature, and is given by:

$$h_{jr} := \begin{cases} 1 \text{ if } r \in \left\{EA^{\mathcal{A}}, NA^{\mathcal{A}}, P^{\mathcal{A}}\right\} \text{ and } j \in \mathcal{A} \\ 1 \text{ if } r \in \left\{EA^{\mathcal{B}}, NA^{\mathcal{B}}, P^{\mathcal{B}}\right\} \text{ and } j \in \mathcal{B} \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

This specification implies that consuming *any premium brand* in the previous period increases one's utility from consuming *any premium brand* in the current period by a magnitude of λ. Such a household is characterized by a "premium" habit in the current period. Similarly, consuming *any generic brand* in the previous period shifts one's utility from consuming *any generic brand* in the current period by λ, a situation we refer to as a "frugal" or "generic" habit.

Our modeling of habit formation is parsimonious in a couple of ways. First, habit is formed toward a class of brands—premium or generic— rather than toward an individual brand. This choice is driven by our motivation: to effectively capture a potentially important mechanism in an emerging market setting characterized by rapid growth in discount brands with minimal advertising.[20] Second, our specification implies that both premium and frugal habits boost household utility by the same magnitude of λ. It is worth noting, however, that a model variant in which we allowed these habits to differ across brand types produced similar magnitudes for the two effects.

The last term in the utility function, ϵ_{ijgt}, represents household and product-specific shocks that follow the Type I Extreme Value distribution and are i.i.d. across households, brands and markets. We complete the utility specification by defining the utility from the outside option, $u_{i\in r, j=0, gt} = \epsilon_{i,0,gt}$. The model's parameters to be estimated are denoted $\theta = \left\{\beta, \alpha, \alpha_{EA}, \alpha_{NA}, \lambda\right\}$. Following familiar terminology from the literature on random-coefficient logit models, we classify these into "linear parameters" $\theta_1 = \left\{\beta, \alpha\right\}$, and "non-linear parameters" $\theta_2 = \left\{\lambda, \alpha_{EA}, \alpha_{NA}\right\}$.

[20] In the appendix we report a robustness check in which we modified our specification to consider brand-loyalty effects. Our findings were qualitatively similar.

We refer to this baseline specification as the *Brand Type Persistence* (BTP) model, and it is on this specification that we base our empirical work. We also estimate a variant of this model which forces λ to equal zero, i.e., it shuts down the habit mechanism. This model variant helps us illustrate the importance of allowing for persistence in household preferences.

Two additional aspects of the demand model are worth noting. First, allowing habit to develop as a consequence of choices made in the preceding period only (as opposed to allowing it to form over a longer history of choices) is consistent with the literature of which we are aware (e.g., Dubé, Hitsch and Rossi 2010). Importantly, the time interval between the current and preceding periods is one or two months. This is longer than the typical interval between shopping trips in the scanner data often used in such applications. In that sense, relying on the preceding period is less restrictive in our application.

Second, consumers in this model are not forward-looking, that is, they make static choices that maximize current-period utility and do not internalize the effect of current choices on future utility. Given the nature of soft drinks (non-durable, relatively inexpensive goods), we view this static modeling approach as appropriate, and, once again, it is consistent with the empirical literature on state-dependent preferences.[21] Nonetheless, the demand patterns implied by this model do reflect a dynamic "persistence" feature, and we seek to understand how this feature interacts with the dynamics in demography (i.e., socioeconomic mobility)—and the implications of those dynamics for competition between premium and generic brands.

The share of type-r households consuming brand j in market gt is given by the logit formula:

$$s_{j,r,gt}(\theta) = \frac{exp(\delta_{jgt} + \alpha_r \cdot p_{jgt} + \lambda \cdot h_{jr})}{1 + \sum_{\ell=1}^{J} exp(\delta_{\ell gt} + \alpha_r \cdot p_{\ell gt} + \lambda \cdot h_{\ell r})}, \tag{4}$$

where $J = 9$ is the number of brands sold in each market. The notation $s_{j,r,gt}(\theta)$ reflects the fact that these shares are model predictions and they depend on parameter values. Brand j's predicted aggregate share is the weighted sum of the shares of the nine household types choosing brand j, where the weights are the population fractions that belong to these types:

$$s_{jgt}(\theta) = \sum_{r \in \mathcal{R}} F_{r,gt} \cdot s_{j,r,gt}(\theta) \tag{5}$$

Consistent with the literature on estimating demand models with discrete-type heterogeneity (e.g., Berry, Carnall and Spiller 1996, Kalouptsidi 2010), the type-specific shares $s_{j,r,gt}(\theta)$ from (4) are not observed in the data, and our estimation strategy is, therefore, based on matching the aggregate shares $s_{jgt}(\theta)$ predicted from (5) with shares s_{jgt} computed from the Nielsen data (see

[21] An alternative is to follow the dynamic estimation literature and model consumers as maximizing an infinite-horizon utility function, making predictions about the future path of prices. Given the nature of the product, we view this as an unnecessary extension.

Section 2). Our framework adapts this approach to the emerging market setup by incorporating into the estimation procedure a dynamic updating mechanism for the fractions $F_{r,gt}$.

Of particular interest are the household-type specific elasticities. The type-r specific own-price ($j = k$) and cross-price ($j \neq k$) elasticities of demand for brand j are computed from:

$$\eta_{jk,r,gt} = \frac{\partial s_{j,r,gt}}{\partial p_{kgt}} \frac{p_{kgt}}{s_{j,r,gt}},$$

where, for brevity, we omit the argument θ, and,

$$\frac{\partial s_{j,r,gt}}{\partial p_{kgt}} = \begin{cases} (\alpha + \alpha_r)\, s_{j,r,gt}\,(1 - s_{j,r,gt}) & \text{if } j = k \\ -(\alpha + \alpha_r)\, s_{j,r,gt} s_{k,r,gt} & \text{if } j \neq k \end{cases}$$

3.2 Dynamic type evolution

Over time, social mobility (as well as rural-to-urban migration) in a particular urban region g changes the aggregate numbers of households in each socioeconomic standing. In addition, in each period t, households make consumption choices that affect the habit state with which they enter period $t + 1$. Both of these processes determine the dynamic evolution of the type-distribution vector $\mathcal{F}_{gt}$ over time. We now fully characterize this dynamic updating process.

We begin by computing $\mathcal{F}_{g1}$, i.e., the type-distribution vector for period $t = 1$ in region g. These values are computed directly from the household-level survey data (HEX 95/96). Recall that this survey was conducted right before our Nielsen data begins, and that it links a household's socioeconomic class to its consumption choice: premium soda, generic soda, or "no soda." Following the discussion in Section 2, for each region in period $t = 1$ we set the number of newly affluent households to zero. By construction, therefore, we set the $t = 1$ population fractions that belong in the three newly affluent types (that is, newly affluent households with premium, generic and "no-soda" habits) to zero. Population fractions at $t = 1$ for the three established-affluent types are set in proportion to the HEX shares for **ABC** households across premium brands, generics, and no soda. Population fractions at $t = 1$ for the three poor types are set analogously using HEX shares for **DE** households.

Given a particular value for the model's parameters θ, these fractions are updated forward for periods $t = 2, ..., 57$. Fixing region g and period t, we explain how to find $\mathcal{F}_{g(t+1)}$ given $\mathcal{F}_{gt}$ and a value for θ. Repeating this updating process for $t = 1, ..., 56$ (and noting that $\mathcal{F}_{g1}$ is known), yields the full trajectory of the distribution of household types over the sample period.

Importantly, a guess of the model's parameters yields a prediction, via (4), of the shares (and masses) of type-r households who consume premium and generic soda in period t. Had we not

allowed social mobility, the computation of $\mathcal{F}_{g(t+1)}$ would be straightforward by simply summing, across the three types in each socioeconomic group, the number of households who in period t consumed a given brand type (premium or generic), and dividing this sum by the period $(t+1)$ total household population. For example, had the newly affluent population been constant over time, the fraction of households who, in time $(t+1)$, are newly affluent and have the premium habit would be computed by predicting the number of newly affluents who consumed premium soda at time t, and dividing by the total population at time t.[22]

The social mobility process, however, complicates these computations. For instance, whenever aggregate upward mobility is detected in a given region between periods t and $(t+1)$, it follows that some of period $(t+1)$'s newly affluent households were poor in period t; to ascertain their $(t+1)$ habit requires information on poor households' choices at time t. Aggregate downward mobility and rural-urban migration between successive periods create similar challenges. To address these challenges and incorporate aggregate data on social mobility and migration to update $\mathcal{F}_{gt}$ into $\mathcal{F}_{g(t+1)}$, we make assumptions on the interaction between social mobility and previous consumption:

Assumption 1 (Socioeconomic Mobility)**.** *Among those households moving up (down) from Poor to Newly Affluent (Newly Affluent to Poor) status, the previous-period shares of premium versus generic brands equal the previous-period shares of premium versus generic brands among all Poor (Newly Affluent) households.*

Assumption 1 implies that social mobility between period t and $(t+1)$ is independent of consumption choices at time t. For example, a household who "moved up" from being poor at t to newly affluent at $(t+1)$ is as likely to have consumed each type of soda at time t as any member in the wider population of poor households at time t. Clearly, other assumptions can be made, and we explain in the appendix that our results are robust to alternative mobility assumptions. The appendix also offers numerical examples of the dynamic updating process implied by our assumptions.

We similarly incorporate an assumption regarding rural-urban mobility, inferred from the observed variation in urban populations since household sizes hardly vary over time, as follows (once again, the appendix demonstrates robustness to this assumption):

Assumption 2 (Migration)**.** *Households moving to urban areas join the Poor socioeconomic group and have a "no-soda" habit. Households moving out of urban areas leave the Poor group, and have premium, generic and no-soda habits in proportion to the shares of those habits among the Poor that remain.*

[22]Recall that there are three types of newly affluent households at time t—those with premium, generic and "no-soda" habits—so the number of newly affluents consuming a premium brand in time t is actually the sum of newly affluents across these three habit states who choose premium soda.

4 Identification and estimation

We now first describe our estimation algorithm, then proceed with intuitive arguments on identification.

4.1 The estimation procedure

The estimation procedure we propose and implement extends the literature on estimating demand models with discrete-type heterogeneity. In that literature, types are often abstract groupings of "similar" consumers, and the population fractions of these types are treated as parameters to be estimated.[23] In our method, in contrast, these population fractions are computed by combining data on aggregate social mobility, and model predictions regarding household choices. The fact that time-t choices determine time-$(t+1)$ habit states requires us to incorporate a dynamic updating routine into each evaluation of the GMM objective function. We now describe the logic of the estimation algorithm, leaving complete technical details to the appendix.

The following steps allow us to construct a GMM objective function and evaluate it at some generic value of the model's parameters θ. Recall that we obtain $\mathcal{F}_{g1}$, i.e., the type-distribution vector for period $t = 1$ in region g, from the HEX data source. Conditional on $\mathcal{F}_{g1}$ and θ, we obtain predictions for aggregate brand shares in period $t = 1$ via equation (5). Using the contraction mapping from Berry, Levinsohn and Pakes (1995), we invert the market share equation that equates the predicted aggregate shares to the shares observed in the data, and solve for the unique vector of base utilities δ for each product that satisfies this equation (noting, as discussed above, that type-specific shares are not observed, and so they cannot be matched).

Using these base utilities, equation (4) now provides us with predictions for type-specific brand choices, from which we can obtain the type-specific fractions of households who, in period $t = 1$, consume premium brands, generic brands, or no soda. This information helps determine habit states for period $t = 2$. Following the discussion in Section 3.2 above, we then use these predicted choices, along with data on region-g's aggregate social mobility between $t = 1$ and $t = 2$, and Assumptions 1 and 2, to compute next period's type-distribution vector $\mathcal{F}_{g2}$.

Repeating this process for periods $t = 2, ..., 56$ (and solving for the base utilities in $t = 57$, the final sample period), and then repeating for each region g, we obtain the base utilities for every brand in every region-period market. From $\xi = \delta - x'\beta - \alpha p$, we can now compute the demand unobservables $\xi_{jgt}(\theta)$ for each brand j in each region-period market gt. The notation reflects the fact that these demand unobservables are computed conditional on particular parameter values.

We follow a familiar approach from the demand estimation literature and make the identifying assumption that these demand unobservables are mean-independent of a set of instrumental

[23] For example, Nair (2007) models video-game consumers as either "high valuation" or "low valuation" types, and estimates their relative population fractions.

variables Z. This assumption gives rise to a GMM objective function that captures the covariance between the instruments and the computed demand unobservables. This approach chooses the parameters which set this covariance as close to zero as possible.

An important feature of this procedure is that inversions of the market share equation cannot be performed independently for different markets over time. One must perform this inversion for period t, obtain the population fractions of each type for period $t+1$, and then perform the inversion for period $t+1$. This dynamic process must be performed at every candidate value of θ considered by the estimation algorithm.

Choice of instruments. We adopt three classes of demand instruments used by Salvo (2009).[24] The correlation of these instruments with prices helps alleviate biases associated with price endogeneity. The first class of instruments are cost shifters, a classic choice of demand instruments. In particular, we use prices of sugar, electricity and fuel.

The second class of demand instruments borrows from Hausman, Leonard and Zona (1994). Specifically, we instrument for a brand's price in a given region with the contemporaneous mean price for this brand in the other regions. The identifying assumption is that prices across different regions are correlated through a common cost structure or through common shifts in the way firms strategically interact (for instance, the mid 1999 premium price cut—see below). This approach can be challenged if common demand unobservables are present (see Bresnahan 1997a, 1997b). However, such issues are of a lesser concern in our setting, for the following two reasons: first, we control for region-specific, brand-level advertising intensity, often absent from demand studies. Second, there is considerable regional variation in demand, explaining the very local nature of Brazilian soft drink distribution and promotion.[25]

Finally, a third set of instruments is afforded by the premium brands' abrupt price cut. We argue that this substantial price cut in mid 1999 was exogenous to the brand-region-time specific demand unobservables ξ_{jgt}. This argument rests on the notion that this large and sudden price drop was plausibly a response to the demographic shifts and expansion of the fringe that we observe over 1996-1999, and not a response to some sudden unobserved mid-1999 demand shock (noting that we also control for advertising intensity, weather shocks and region-specific drifts). In practice, we generate a dummy variable which takes on the value 1 for all time periods after July 1999, and interact it with brand-region fixed effects, thus allowing the effects of this supply-side shift to vary by brand within each region.

4.2 Identification

This section provides intuitive arguments for identification of our model (beyond overcoming

[24] Salvo (2009) estimates an AIDS demand model, a different approach compared to the discrete-choice model we offer in this paper. Just the same, instrumenting for price endogeneity is similarly relevant to both frameworks.

[25] In this context, it is worth noting that the penetration of national retailers in Brazil is still limited relative to the United States.

price endogeneity, as discussed above). In particular, we explain what variation in the data is helpful for identifying our habit mechanism, as well as the heterogeneous price sensitivities across different socioeconomic groups. While the literature often identifies state-dependent preferences from household-level panels of survey data, we explain how the emerging market setup enables identification using rich cross-sectional and temporal variation in *aggregate* data on market shares, prices and social mobility. We emphasize two kinds of data variation: socioeconomic transitions and price changes.

Socioeconomic transitions. Shifting demographics play a crucial role in our identification strategy. We observe both the growth of the middle class from 1996/97 on, and the subsequent partial reversion during the recession that started around 2000/01. Importantly, these transitions occurred at differential rates across regions. While our inclusion of brand-region fixed effects controls for *fixed* differences in preferences across regions—stemming, for instance, from cultural or historical reasons—the intra-region temporal variation provides a key source of identification.

To illustrate this point, consider two regions that vary substantially in terms of their dynamic evolution: region 1 (the Northeast) and region 4 (São Paulo Metro). The following tables depict some socioeconomic and product market data for these regions at several points in time:

% Urban	Dec/Jan 1997		Dec/Jan 2000		Dec/Jan 2003	
households	Poor	New. Affl.	Poor	New. Affl.	Poor	New. Affl.
Region 1 (Northeast)	65%	–	44%	24%	57%	15%
Region 4 (São Paulo Metro)	36%	–	23%	16%	23%	19%
Shares: No Soda	Dec/Jan 1997		Dec/Jan 2000		Dec/Jan 2003	
& Generics	s_0	s_{gen}	s_0	s_{gen}	s_0	s_{gen}
Region 1 (Northeast)	87%	0.3%	82%	5.0%	80%	8.4%
Region 4 (São Paulo Metro)	61%	6.6%	62%	12.6%	59%	11.9%

Sources: Nielsen, IBOPE, IBGE (PNAD). Market shares s_0 and s_{gen} are for the outside option and for generics, respectively.

At the start of our sample, region 1 is substantially poorer than region 4 (65% of region 1's urban households are poor vis-à-vis 36% for region 4) and, at the same time, exhibits lower soda penetration relative to its wealthier counterpart (87% of region 1's households do not consume soda against 61% for region 4). Notice that this cross-sectional variation can in principle be explained not only by the poor being more price sensitive than the established affluent, but also by region 1 potentially exhibiting a lower preference for soda relative to region 4. Such fixed differences, however, are controlled for with brand-region fixed effects in the utility specification.

From 1997 to 2000, region 1 boasted stronger upward mobility relative to region 4: by 2000, 24% of region 1's households were newly affluent compared with 16% of region 4's households.

Over these same years, region 1's soda penetration $(1 - s_0)$ grew substantially, from 13% to 18%, while soda penetration in region 4 was about flat. Generic brands in region 1 enjoyed a huge gain in share (s_{gen}) from 0.3% to 5.0%, while in region 4, the share of generics "only" doubled.

This joint temporal variation in social mobility *and* soda consumption choices helps us identify price sensitivity parameters. Notice that our model allows price sensitivity to vary by socioeconomic standing, i.e., the price sensitivities of the poor, newly affluent and established affluent are given by α, $(\alpha + \alpha_{NA})$ and $(\alpha + \alpha_{EA})$, respectively. Intra-regional social mobility of households between poor and newly affluent status thus changes the aggregate price sensitivity in the region. The co-variation of this price sensitivity with aggregate market shares, controlling for prices, identifies the price sensitivity parameters.

The tables above also demonstrate the differential effects of the recessionary period in the two regions. Region 1 saw the proportion of newly affluent households shrink considerably, from 24% in 2000 to 15% by 2003, yet the penetration of soda consumption continued rising, even if at a lower rate, from 18% to 20%. Importantly, the recession was not accompanied by declining soda prices (see Figure 2). This pattern is suggestive of persistence in preferences. Such data variation is quite effective in identifying the parameter λ. It demonstrates how persistence manifests itself directly in our data, as opposed to being an artifact of an econometric strategy. We provide another example of such variation below.

The stability of soda consumption in the recessionary period can be contrasted with the substantial decline in the sales of cement during those years, as depicted in Figure 3, following years of common growth. This differential pattern is suggestive of stronger persistence in preferences over food and beverage, such as soft drinks, compared to other product categories.

We further argue that controlling for persistence in preferences actually helps us identify the price sensitivity parameters. To see this, imagine that we did not allow for such a persistence feature. Our model would then interpret the data variation in the recessionary period as evidence that the poor are "not that price sensitive," since all one would see through the lens of such a model is an increasingly poor population consuming stable amounts of soda. It would then be difficult to elicit greater price sensitivity among poor households compared to the more affluent groups. This intuition is consistent with estimation results for a model variant in which we shut down the habit mechanism, as we report below. Controlling for persistence, therefore, is not only important in its own right, but plays a key role in identifying other dimensions of household preferences. To our knowledge, this is the first paper that makes this point. It provides another important reason to account for habit formation in the study of demand in emerging markets.

Price variation. Another important source of identification stems from price variation which can be viewed as exogenous to unobservable demand shocks. First, consider the gradual but relentless price reduction in the competitive fringe over the period 1996 to 2000, as captured in

Figure 2. The close substitutability among those generic brands suggests that their price should, to a large extent, stay close to their marginal cost of production and distribution. This suggests that the decline in fringe prices was predominantly driven by supply side (cost) considerations. Expanded capacity and scale, learning effects and exit of inefficient producers may all have contributed to declining fringe costs and prices.

Through most of the period of declining fringe prices, generics were able to grow substantially at the expense of the premium brands and the outside good. To stay with the examples from the data (see the table of shares above), between 1997 and 2000: (i) in region 1, the generic share grew by 5 points, with the premium share (i.e., $1 - s_{gen} - s_0$) holding up quite well; and (ii) in region 4, the generic share grew by 6 points, and at the expense of the premium share. Intuitively, the co-movement in prices and shares is picked up by the parameters that govern price sensitivity and habit formation, thus contributing to their identification.

Consider also the abrupt premium price cut in mid 1999. As argued in subsection 4.1, this decision can be viewed as largely exogenous to the demand unobservables ξ_{jgt}. Notice that the generic share held up quite well following this premium price cut. This can be seen in the tables above for regions 1 and 4, and in Figure 2 for all regions combined. We view this as further evidence of persistence in preferences. Specifically, it supports our Brand Type Persistence (BTP) mechanism: a habit of "going generic," developed by part of the population prior to the mid 1999 premium price cut, made it very difficult for premium brands to win such households over, even via a drastic price cut. The price cut did, however, help protect premium brands from *further* market share losses, in part by attracting households who, at the time, had a "no-soda" habit. We return to this discussion in the results section.

To sum, largely exogenous variation in both prices and in the socioeconomic composition of households provides us with means to identify both heterogeneous price sensitivities and our habit mechanism. That said, we do treat prices, in general, as endogenous, as explained above.

5 Results

We start by reporting estimates obtained from our baseline demand model and comparing these to estimates from a "no habit" model variant. We subsequently employ our estimates to examine, via a counterfactual analysis, the premium sellers' strategic price cut in mid 1999.

5.1 Estimates from the demand model

Table 3 reports estimates for our demand model. The price sensitivity of the poor socioeconomic group, α, has the expected negative sign and is very precisely estimated. The parameters α_{NA} and α_{EA} are also precisely estimated and are positively signed. Recalling that $(\alpha+\alpha_{NA})$ and $(\alpha+\alpha_{EA})$

capture the price sensitivities of the newly affluent and the established affluent, respectively, we obtain, intuitively, that the affluent groups are less price sensitive than the poor. Also note that both $(\alpha + \alpha_{NA})$ and $(\alpha + \alpha_{EA})$ are estimated to be negative.

Are newly affluent households more price sensitive than established affluent households? As discussed in the introduction, such a finding could help explain the success of generic brands, as well as Coca-Cola's deep price cut. Our estimates, however, do not lend support for such a claim. While the point estimates do reflect that $\hat{\alpha}_{NA} < \hat{\alpha}_{EA}$, the difference is not statistically significant, and is small in terms of economic significance, as we demonstrate below with an analysis of demand elasticities. *We do not, therefore, find evidence that the emerging middle class is more price sensitive than the established middle class.*[26]

In contrast, our findings do provide strong evidence for the second mechanism we study in this paper: habit formation. The coefficient λ is estimated to be positive and it is very precisely estimated.[27] To provide a sense of economic significance, notice that $\lambda/|\,(\alpha + \alpha_{NA})\,|$ measures the increase in the willingness to pay of a newly affluent household for a liter of generic (premium) soda resulting from previous-period consumption of generic (premium) soda. The implied increase is $5.09/|(-5.25 + 2.75)|$, or R$ 2.04. Further, the implied increase in willingness to pay for a generic over a premium brand when the newly affluent household has a generic habit *rather than a premium habit* is twice this amount.

These measures indicate a substantial monetary value of habit formation, and a crucial role played by this mechanism in emerging market dynamics. Once a newly affluent household develops a generic habit, "convincing" it to switch to a premium product becomes substantially more difficult. This helps explain the sense of urgency with which premium brands acted in mid 1999, as we argue in the counterfactual analysis below.

Table 3 further reports the effects of several shifters of δ_{jgt}, the base utility of consuming brand j in region-period market gt. We control for $9 \times 7 = 63$ brand-region fixed effects, capturing the tendencies of particular regions to consume different brands (e.g., historically, tastes for Pepsi are known to be relatively strong in the South, region 6—see Table 1). We also control for bi-monthly seasonality effects interacted with brand type (i.e., premium versus generic), to allow these effects to differ across product types. Over and above seasonality, market gt's mean temperature has a positive and significant effect on demand.[28]

[26] The appendix reports robustness checks. The finding that newly affluents are not significantly more price sensitive than established affluents holds rather consistently across the bulk of the different specifications we tried.

[27] One possible interpretation for persistence in market shares could be serial correlation in the demand shocks ξ. Our inclusion of brand-region fixed effects and region-specific trends makes this possibility less of a concern. Just the same, we have calculated the simple correlation between current-period and previous-period estimated values for ξ for each brand-region combination, yielding 63 such correlations. Most of these correlations are small in absolute value and many of them are negative. In a handful of brand-region combinations, positive correlations as high as 0.5 to 0.7 are observed. Most of those are in region 1, and they do not pertain to the leading brands such as Coke and Diet Coke.

[28] To illustrate within-season variation in temperature, winter temperatures in the southern region 6 averaged 15.1^oC in July 2001 against 12.1^oC in July 2000.

Our specification also includes region-specific effects of brand-level media advertising.[29] We interact advertising GRP with regional dummy variables, to allow advertising effects to differ across regions (reflecting, for instance, varying levels of ownership of household electronics and exposure to media, and the fact that these measures pertain to only the main cities within each region). The effects of advertising intensity are positive in all seven regions, although in most cases not statistically significant. Finally, coefficients on region-specific time trends are negative and mostly precisely estimated. Such negative effects are consistent with continued improvement in the value of the outside option, which includes beverages other than soda such as juices and (tap or bottled) water, concomitant with the overall trend of economic growth.[30] The trend variables are rescaled to vary from 0 at the start of the sample period to 1 at its end, and thus the effects reported in the table are economically small.

To further explore the economic implications of our demand estimates, Table 4 reports price elasticities. The table lists both aggregate own-price elasticities for the leading brands, and own-price elasticities by household type, computed as means over all region-period markets gt. A 1% increase in Coke's price lowers its market share by 1.7%, compared with somewhat larger (in magnitude) elasticities of -2.1 for the other premium brands, Guaraná Antarctica, Fanta and Pepsi. The own-price elasticity for generics is -0.5. While this value may seem low, note that this is the elasticity of demand for the aggregation of generic brands. The demand for each individual generic brand should be much more elastic, given the limited differentiation and fierce price competition within the fringe.

Examining the nine type-specific elasticities for Coke, and fixing the habit state, we see that demand becomes more elastic the lower is the socioeconomic standing. For instance, considering households with a premium habit, the elasticities are (-1.5,-1.7,-5.0) for the established affluent, newly affluent, and poor groups, respectively. The demand elasticity of the newly affluent is much closer to that of the established affluent than to that of the poor. Further to the discussion above, we find that the difference in price sensitivity between the new and the established middle class is not significant either statistically or economically.

Fixing the socioeconomic standing, the habit state has a strong impact on demand elasticities. Considering, for example, the newly affluent group, demand for Coke is least elastic for households with a premium habit (-1.7). Households with "competing habits"—either generic or "no-soda"—exhibit higher elasticities of demand for Coke (-2.7 and -2.6, respectively).

Further illustrating these findings, Figure 5 plots the evolution of own-price elasticities, by household type, for the Coke brand in region 4. Demand by all groups becomes less elastic

[29] To gain a sense of variation in such measures, the advertising intensity for the Coke brand in São Paulo Metro (region 4) amounted to 2199 Gross Rating Points in December 2000, rising to 3587 GRP in December 2001, while Pepsi's GRP were 351 and 598 respectively in these same periods.

[30] Forbes (2004) reports that the "the juice category grew twenty times over the past decade, albeit from a low base." IBGE's annual household surveys (PNAD) also indicate a sustained increase in access to tap water and piped sewerage in urban Brazil.

halfway through the sample, when premium brands cut prices. The figure separates the nine types into three distinct groups; the least elastic demand for Coke is by established affluent *and* newly affluent households with a premium habit. The most elastic demand is by the three poor types, with the other four types (established affluent and newly affluent with generic or no-soda habits) displaying intermediate price sensitivities. This picture further underscores our point that habit plays a crucial role: conditional on a premium habit, it is hard to tell the difference between the established affluent and the newly affluent, whereas the demand by affluent households with other habits is much more elastic.

Table 5 reports predictions of type-specific consumption choices from the estimated demand model. Overall soda penetration (i.e., the share of households who purchase soda) is 51%, 37% and 3% for the established affluent, the newly affluent and the poor, respectively (these are means computed over all region-period markets). More affluent households are also more likely to favor premium over generic brands: the premium-to-generic consumption ratios are 2.0, 1.5 and 0.6 for the established affluent, newly affluent and poor groups, respectively.

A "no habit" model variant. Table 6 presents results from a variant of our demand model which shuts down the habit formation mechanism. To be clear, this model is identical to the baseline model in Table 3 except that the parameter λ is constrained to equal zero. Estimates from this model still suggest that established affluent households are less price sensitive than the poor ($\hat{\alpha}_{EA}$ is positive and statistically significantly different from zero) but the price sensitivity difference across these two groups narrows compared to the estimates from the baseline model.

Importantly, the "no habit" model variant does not suggest that the newly affluent are less price sensitive than the poor ($\hat{\alpha}_{NA}$ is negative and statistically insignificant). This result stands in contrast to our baseline model. The fact that the no-habit model variant does not separate the price sensitivities of the newly affluent from that of the poor is consistent with arguments provided in the identification section above: during the recessionary period in the later part of the sample, households moved down from newly affluent to poor status, yet soda consumption remained stable. By not allowing a habit mechanism, the model must interpret this as evidence that the price sensitivity of these two groups is similar.

We view the predictions of the baseline model as more realistic that those of the "no habit" model variant. It is well-documented that Brazil's emerging middle class exhibited lower price sensitivity than the poor, given the demand surge observed across many consumer goods markets, soft drinks being one of them. The predictions of the no-habit model seem to suggest that the emergence of a new middle class did nothing to change the aggregate price sensitivity, since it predicts that the newly affluent are as price sensitive as the poor. This model variant, therefore, entirely misses the phenomenon which is at the heart of our study: an expansion in demand stemming from a socioeconomic transformation.

5.2 Counterfactual analysis of the premium price cut

One of the striking features of the data is the premium brands' sharp price cut, led by Coca-Cola, almost halfway through the sample period. As the solid lines in the left panel of Figure 6 indicate for region 5 (this variation is similar for other regions), per-liter premium brand prices stayed broadly flat at about R$ 1.15 until mid 1999, then dropped—abruptly—to R$ 0.90 and stayed at this lower level. Fringe prices, in contrast, experienced a prolonged, gradual decline from R$ 0.80 to R$ 0.55 between late 1996 and mid 2000. The picture reveals that fringe prices did not deviate from their downward trend in response to the premium price cut, consistent with fringe prices closely tracking their producers' marginal costs.

We employ the estimated model to simulate the evolution of market shares had premium sellers *not* cut prices in mid 1999. This counterfactual price path is marked by the dashed line in the left panel of Figure 6. In this analysis, we keep fringe prices equal to the ones observed in the data. This assumption is justified by the fringe's competitive nature and, as discussed, the absence of an apparent pricing response to the premium price cut.

The right panel reports the estimated impact on aggregate premium and generic market shares. Observed shares are marked by solid lines, whereas counterfactual shares are marked by dashed lines (shares in this figure are out of the total market size, which includes the outside option, so that the premium and generic shares do not sum to one). A clear picture emerges: had premium producers failed to cut prices, they would have suffered a deep and substantial market share loss, hitting a rock bottom in the winter of 2000. At that point, the counterfactual premium market share would have been 12%, compared to a share of over 20% in the observed sample.

The counterfactual scenario is marked by the relentless growth of generic brands at the expense of their premium competitors. The analysis suggests that the generic market share would have surpassed the premium share early in 2000. By 2003, generics in region 5 would have enjoyed a market share advantage over premium brands of 10% (24% against 14%).

This analysis provides support for Coca-Cola's price cut, in that it seems to have prevented a substantial drop in market share. An important insight from the analysis is that the premium price cut was especially effective in terms of attracting customers who otherwise would have chosen the outside, "no soda" option. It was less effective in terms of converting consumers of generic brands into premium consumption. For example, inside shares at actual (reduced premium brand) prices over 2001-02 average 46% (28% premium plus 18% generic) to be compared with inside shares of 38% (16% premium plus 22% generic) at (higher) counterfactual prices. This is suggestive of substantial market segmentation, consistent with the habit mechanism that limits the scope for "business stealing" across the types of brand offerings, and with the high monetary value of habit formation. Still, the 4 percentage point growth in the generic share would have represented almost a one-quarter increase (+4.2/17.6) in the fringe's penetration.

Impact on variable profit. While the analysis above suggests that Coca-Cola's price cut succeeded in avoiding a deep market share loss, we note that this was achieved at a cost: a deep price cut of over 20%. In other words, premium sellers sacrificed a non-negligible portion of their margins to protect their market shares.[31] To assess the overall impact of the price cut on earnings, we perform a back-of-the-envelope calculation of variable profit both in the observed sample, and under the counterfactual (no price cut) scenario.

Using information gleaned from Ambev's local SEC filings, conversation with industry insiders, among other sources, we estimate that the premium brands' combined variable profits (excluding fixed costs) during the first three years after the price drop amounted to R$ 860 million, to be compared to counterfactual profits of R$ 740 million, had the price cut not occurred.[32] That is, a 14% loss in variable profit over the medium run was avoided by the premium price cut. The evidence supports the notion that the price cut was beneficial in terms of its impact on both market shares and profits.

A comparison with the "no habit" model variant. We wish to explore the role played by the habit mechanism in this analysis. To this end, Figure 7 performs the same counterfactual analysis but using estimates from the no-habit model variant discussed above. It is clear from comparing Figure 7 to Figure 6 that the no-habit model is associated with a substantially smaller erosion of market share for premium brands had they failed to cut prices. Further, using the same back-of-the-envelope calculations discussed above, the no-habit model implies that the price cut actually *decreased* premium brands' variable profit, from R$ 1.0 billion (with no price cut) to R$ 860 million (with price cut) over the same three years. This stands in stark contrast to the predictions of the baseline model.

Discussion. Though examining the premium brands' pricing policy is the subject of a sequel paper, the counterfactual analysis lends strong justification for Coca-Cola's strategic price cut. Importantly, this conclusion is delivered by the baseline model, but not by the no-habit model variant, highlighting the role played by habit formation in this emerging market. Our discussion of the estimation results above suggested that habit carries a large monetary value. In particular, once a newly affluent household "goes generic," it is significantly less likely to switch into consumption of premium, expensive soda. The counterfactual analysis demonstrated how this feature can wreak havoc on the market share of premium brands: had they not cut prices in mid 1999, the generic fringe would have continued to gain ground, while premium brands would have lost considerable market shares and profit.

Our analysis shows that Coca-Cola's price did not allow it to convert many households with the generic habit into consumption of its premium products. Rather, the main effect was to tap

[31] Protecting market shares, even at a high cost, may be rational insofar as current market share is an "asset," predictive of future profit. See Bronnenberg, Dhar and Dubé (2009) on the persistence of brand market shares.

[32] See the appendix for details on how this calculation was performed.

into the large pool of households with the no-soda habit. By inducing a substantial portion of these households to develop a habit of consuming premium soda, Coca-Cola and Ambev were able to shield themselves against further losses. Our analysis suggests that it is this mechanism that stabilized market shares after mid 1999, as demonstrated in Figure 2.

Our ability to draw such conclusions stems from the richness of our framework, which captures both social mobility and habit formation. In contrast, the more standard "no habit" model variant fails to separately identify the price sensitivities of the newly affluent and the poor. Moreover, it misses the crucial role played by persistent preferences in the dynamics of competition between premium and generic brands in a rapidly changing market.

6 Concluding remarks

This paper examines two salient features of the Brazilian soft drink market: the emergence of a new middle class, and the rapid growth of a generic fringe. Using unique data with very rich cross-sectional as well as temporal variation, we estimate a model that highlights two aspects which we view as highly important in such markets: the heterogeneous price sensitivities of different socioeconomic groups, and habit formation in household preferences.

Our brand type persistence (BTP) mechanism captures a world in which premium brands are prompted to cut prices in the wake of an emerging middle class. If they fail to do this, a substantial mass of the "new customers" might be captivated by the generic habit. It may then prove much more difficult to convince these consumers to pay substantially more for a highly advertised premium brand.

While our application focuses on the Brazilian soft drink market, we view the issues tackled in this work as highly pertinent to many consumer goods markets in the developing world, where a tension between advertised branded offerings and discounted generics exists or is developing. Understanding the features of demand and the microeconomics of competition in such markets should be of great interest for policymakers and firms alike.

References

[1] ABEP, Brazilian Association of Market Research Firms (2003). Critério de classificação econômica Brasil: Sistema de pontos. ABEP Publication, São Paulo. (Points system for classifying households into economic segments: Brazil criteria, in Portuguese.)

[2] Ambev (2003). Informações anuais (IAN). Comissão de Valores Mobiliários, CVM. (Annual report to Brazil's Securities and Exchange Commission, in Portuguese.)

[3] Atkin, D. (2011). Trade, tastes and nutrition in India. Mimeo, Yale University

[4] Bils, M. and P. J. Klenow (2004). Some evidence on the importance of sticky prices. *Journal of Political Economy* 112, 947-985

[5] Boston Consulting Group (2002). Mercados pouco explorados: Descobrindo a classe C. BCG Publication, São Paulo. (Under-explored markets: Discovering the "C" class, in Portuguese.)

[6] Berry, S., J. Levinsohn, and A. Pakes (1995). Automobile prices in market equilibrium. *Econometrica* 63, 841-90

[7] Berry, S., M. Carnall, and P. T. Spiller (1996). Airline hubs: Costs, markups and the implications of consumer heterogeneity. NBER Working Paper 5561

[8] Bresnahan, T. F. (1997a). Valuation of new goods under perfect and imperfect competition: A comment, in Bresnahan, T. F. and R. Gordon (eds.), *The Economics of New Goods, Studies in Income and Wealth* Vol. 58. National Bureau of Economic Research, Chicago

[9] Bresnahan, T. F. (1997b). The Apple Cinamon Cheerios war: Valuing new goods, identifying market power, and economic measurement. Mimeo, Stanford University

[10] Bronnenberg. B., S. Dhar and J.-P. Dubé (2009). Brand history, geography, and the persistence of brand shares. *Journal of Political Economy* 117, 87-115

[11] Bronnenberg. B., J.-P. Dubé and M. Gentzkow (2012). The evolution of brand preferences: Evidence from consumer migration. *American Economic Review*, forthcoming

[12] Calzolari, G., A. Ichino, F. Manaresi, and V. Nellas (2012). When the baby cries at night: Uninformed and hurried buyers in non-competitive markets. Mimeo, University of Bologna and Bank of Italy

[13] Carvalho Filho, I. and M. Chamon (2011). The myth of post-reform income stagnation: Evidence from Brazil and Mexico. *Journal of Development Economics*, in press

[14] Chaudhuri, S., P. K. Goldberg and P. Jia (2006). Estimating the effects of global patent protection in pharmaceuticals: A case study of quinolones in India. *American Economic Review* 96, 1477-1514

[15] Dubé, J.-P., G. J. Hitsch, P. E. Rossi and M. A. Vitorino (2008). Category pricing with state-dependent utility. *Marketing Science* 27, 417-429

[16] Dubé, J.-P., G. J. Hitsch and P. E. Rossi (2010). State dependence and alternative explanations for consumer inertia. *RAND Journal of Economics* 41, 417-445

[17] Economist (2007). Adiós to poverty, hola to consumption: Faster growth, low inflation, expanding credit and liberal trade are helping to create a new middle class in Latin America. Print edition of August 18, 2007

[18] Economist (2011a). Politics in emerging markets: The new middle classes rise up. Print edition of September 3, 2011

[19] Economist (2011b). Indonesia's middle class: Missing BRIC in the wall. Print edition of

July 21, 2011

[20] Economist (2012). China's motor industry: Stepping on the gas. Edition of April 24, 2012

[21] Eichenbaum, M. S., L. P. Hansen and K. J. Singleton (1988). A time series analysis of representative agent models of consumption and leisure choice under uncertainty. *The Quarterly Journal of Economics* 103, 51-78

[22] Erdem, T. (1996). A dynamic analysis of market structure based on panel data. *Marketing Science* 15, 359-78

[23] Financial Times (1999). Brazil's regional drinks makers slake thirst for value: The tax regime and growing demand have penalised leading brands. Edition of June 16, 1999

[24] Forbes (2004). Segmento de sucos cresceu 20 vezes nos últimos dez anos, mas brasileiro ainda consome pouco. (Juice category grew twenty times over the past decade, yet Brazilians still consume little, in Portuguese.) Edition of February, 2004

[25] Frankel, D. M. and E. D. Gould (2001). The retail price of inequality. *Journal of Urban Economics* 49, 219-239

[26] Galanz (2008). Galanz household electric appliances: the unruly competition in the market will hopefully end. 'Corporate News' posted on the company's website, dated September 1, 2008

[27] Hausman, J. A., G. K. Leonard and J. D. Zona (1994). Competitive analysis with differentiated products. *Annales D'Economie et de Statistique* 34, 159-80

[28] Hurwitz, M. A. and R. E. Caves (1988). Persuasion or information? Promotion and the shares of brand name and generic pharmaceuticals. *Journal of Law and Economics* 31, 299-320

[29] Kalouptsidi, M. (2010). From market shares to consumer types: Duality in differentiated product demand estimation. *Journal of Applied Econometrics*.

[30] Keane, M. P. (1997). Modeling heterogeneity and state dependence in consumer choice behavior. *Journal of Business & Economic Statistics* 15, 310-27

[31] Lach, S. (2007). Immigration and prices. *Journal of Political Economy* 115, 548-587

[32] Nair, H. (2007). Intertemporal price discrimination with forward-looking consumers: Application to the US market for console video-games. *Quantitative Marketing and Economics* 5, 239-92

[33] Neri, M. (1995). Sobre a mensuração dos salários reais em alta inflação. (About the measurement of real wages under high inflation, in Portuguese.) *Pesquisa e Planejamento Econômico* 25, 497-525

[34] Nevo, A. (2000). A practitioner's guide to estimation of random coefficients logit models of demand. *Journal of Economics & Management Strategy* 9, 513-548

[35] Nevo, A. and Konstantinos Hatzitaskos (2006). Why does the average price paid fall dur-

ing high demand periods? Mimeo, Northwestern University and University of California, Berkeley

[36] Salvo, A. (2009). Cut-throat fringe competition in an emerging country market: Tax evasion or the absence of market power? *Journal of Industrial Economics* 57, 677-711

[37] Salvo, A. (2010). Inferring market power under the threat of entry: The case of the Brazilian cement industry. *RAND Journal of Economics* 41, 326-350

[38] Sancheti, S. and K. Sudhir (2009). Education consumption in an emerging market. Mimeo, Yale University

[39] Scott Morton, F. M. (2000). Barriers to entry, brand advertising, and generic entry in the US pharmaceutical industry. *International Journal of Industrial Organization* 18, 1085-1104

[40] SuperHiper (2003). Survey presents consumer purchasing habits at retail. Trade publication, edition of 03/2003

[41] Wall Street Journal (2011). A rags-to-riches career highlights Latin resurgence. Front page article in the print edition of November 15, 2011

FOR ONLINE PUBLICATION

A Appendix

A.1 Data sources

We refer the reader to Salvo (2009) for further details on Nielsen's soda market panel and McCann-Erickson's media advertising intensity panel. We use regional average monthly temperatures made available by the National Institute of Meteorology (INMET). Wholesale prices of refined sugar (the 'IPA-OG açúcar') and a transport fuel price index (the 'IPA-OG combustíveis e lubrificantes') prepared by the Fundação Getúlio Vargas were obtained from the Institute for Applied Economic Research (IPEA), and regional high-voltage electricity prices ('classe industrial') were provided by the National Agency for Electrical Energy (ANEEL). As with soft drink prices, we inflation-adjust nominal factor prices using a consumer price index (the 'IPC-br') published by the Fundação Getúlio Vargas. The CPI has averaged +7.8% per year over the sample period.

In what follows, we explain how we combine IBOPE's LatinPanel survey with IBGE's annual household surveys (the PNAD, 'Pesquisa Nacional por Amostra de Domicílios') to produce household counts by socioeconomic standing. We then describe how we obtain the type-distribution vector for the first period, $\mathcal{F}_{g1}$, from IBGE's 1995/96 urban household expenditure survey (the POF, 'Pesquisa de Orçamentos Familiares').

Data on aggregate social mobility. From IBOPE's LatinPanel we observe the proportion of urban households that belong in either the **ABC** or **DE** socioeconomic groups by year time

(see the text) and across regions. IBOPE's regions map directly into Nielsen's 7 regions, with the exception of region 1—states in the Northeast excluding Maranhão and Piauí—for which IBOPE's coverage includes *all* states in the Northeast as well as states in the North. Since Maranhão, Piauí and the North comprise the country's least urbanized and least populated area, we simply take IBOPE's urban distributions for the Northeast/North as representative of urban households in Nielsen's region 1. IBOPE's survey through 2002 was representative of all municipalities with populations of at least 20,000, and their coverage was expanded in 2003 to represent municipalities with populations exceeding 10,000.

We obtain household counts from IBGE's annual household surveys (PNAD). These cover households, both urban and rural, in all 27 states of the country. For perspective, 115,654 households were sampled in 1999. IBGE's household-level weights allow us to expand the representative sample to the universe of households. We consider only households residing in urban areas and in states within each Nielsen region. For example, for region 1, we sum the number of urban households across all states in the Northeast less Maranhão and Piauí.

We then multiply, for each Nielsen region and year, the IBOPE socioeconomic proportions of urban households by the IBGE urban household counts. To increase the frequency of the resulting panel from annual to monthly periods (or bimonthly periods, thus matching the frequency of Nielsen's point-of-sale audits), we linearly interpolate from September of one year to September of the following year, noting that September is the IBGE PNAD's annual "month of reference."

Data on household-level brand choices. IBGE's HEX (POF) 95/96 surveyed 16,013 households in 11 large metropolitan areas across the country. Carvalho Filho and Chamon (2011) discuss this survey in detail. Over a reference period of one week falling between October 1995 and September 1996, the soft drink expenditure in R$ for consumption inside the home is recorded for each household, detailed by soda brand(s) purchased. We then classified the following brand descriptions and codes as "premium" brands: *Coca-Cola* (9301), *Pepsi* (9302), *Guaraná* (9303), *Fanta laranja, uva, limão* (9304), *Soda limonada* (9307), *Mirinda* (9308), *Sukita* (9315), *Pop laranja* (9316), and *Refrigerante água tônica* (9349). Examples of coded brand descriptions that we classified as "generic" brands are: *Refrigerante tubaína* (9318), *Refrigerante laranja exceto Fanta, Sukita, Pop, Crush* (9339), *Refrigerante cola exceto Coca-Cola e Pepsi-Cola* (9340), *Refrigerante cajú qualquer marca* (9346), and *Refrigerante Goianinha* (9355). Of the 16,013 households, 10,172 (or 64% of households) were recorded as making no soda purchases, 4,465 households (28%) purchased only brands that we can confidently identify as premium, 310 households (2%) purchased only brands that we identify as generic, and 236 households (only 1%) simultaneously purchased brands that we identify as premium and brands that we identify as generic. This observation justifies our modeling of soda-consuming households at each point in time as either premium or generic shoppers, but not "hybrids."

We deemed four soda descriptions to be ambiguous with regard to brand type: *Refrigerante água natural* (9310), *Refrigerante gasosa* (9319), *Refrigerantes não especificado* (9335) and *Refrigerante dietético* (9360). We need to assign the soda expenditure of the remaining 830 soda-purchasing households (5% of the survey sample) to either premium brands or generic brands. These are households whose soda expenditure we cannot entirely identify by brand type, such as a household purchasing R$ 4 of Coke (9301) and R$ 2 "Soda not specified" (9335). To do this, we first designate as premium the "brand-unidentifiable" soda expenditure portion (R$ 2 in the example) for those households whose identifiable-premium expenditure share of soda exceeds 50% (Coke's 67% share in the example) *and* identifiable-generic expenditure share of soda is less than 10% (0% in the example). Similarly, we assign to generic the unidentifiable soda expenditure portion for those identified-generic-dominant households. Finally, the soda expenditure portion that for a remaining 614 households is still not assigned to a brand type—e.g., a yet to be assigned R$ 2 of "Soda not specified" (9335) purchased by another household—is allocated among premium and generic expenditure: (i) in proportion to the (identified or designated) premium versus generic expenditure shares within the household; or (ii) should none of the household's soda expenditure be identifiable (e.g., a household who purchased R$ 2 of "Soda not specified" (9335) only), the allocation is done in proportion to the premium versus generic expenditure shares across households in the same socioeconomic group and metropolitan area. (We use balance sheet data to classify households according to socioeconomic standing, as described in Section 2, and use IBGE's weights to expand the representative sample to a universe of 12.5 million households across the 11 metropolitan areas.)

To calculate household-level premium and generic quantities, we divide HEX 95/96 expenditures on premium and generic soda by Nielsen's region-specific share-weighted mean prices for premium and generic brands, respectively, on $t = 1$ (December 1996-January 1997). We then aggregate premium (resp., generic) quantities across the universe of households belonging to each socioeconomic segment (**ABC** or **DE**) living in the HEX-surveyed metropolitan areas for each Nielsen region g (e.g., the cities of Recife, Fortaleza and Salvador in the Northeast, $g = 1$). The premium (resp., generic) soda shares among the initial masses of established affluent and poor households are calculated analogously to how we define $s_{jgt} = q_{jgt}/\mathcal{M}_{gt}$ in Section 2, i.e., taking market size (in our base specification) as six liters per household per week times the number of weeks in period $t = 1$. Combining these premium versus generic (versus no soda) shares by socioeconomic group with first-period household counts by socioeconomic group (as per above), yields $F_{EA^{\mathcal{A}},g1}, F_{P^{\mathcal{A}},g1}$ and $F_{EA^{\mathcal{B}},g1}, F_{P^{\mathcal{B}},g1}$ (recall that $F_{NA^{\mathcal{A}},g1} = F_{NA^{\mathcal{B}},g1} = 0$). In our base specification, we consider soda purchases recorded as being for the household's inside-the-home consumption (rather than "individual consumption") and at stores coded as *Supermercado* (1), *Hipermercado* (2), *Padaria* (3), *Lanchonete* (11), and *Mercado & Central de Abastecimento* (26),

in view of the mapping to Nielsen's self-service channel (stores with checkouts). Further details can be provided upon request.

Two points are noteworthy. First, the HEX survey suggests that household size does not vary significantly across socioeconomic group. The mean size across **ABC** and **DE** households—all urban—is 3.64 (std. dev. 1.58 across 7,916 households) and 3.76 (s.d. 1.99 across 8,097 households), respectively. Second, while the HEX shares that enter the initial conditions $\mathcal{F}_{g1}$ are calculated following the market share definition of Section 2, one should note that the shares reported in Table 2 are extensive margins of soda consumption, i.e., the proportions of households who purchase any quantity of soda for the home (and which type). For brevity, we do not report "intensive margins"—intensity of soda consumption conditional on positive consumption—but figures are available from the authors upon request. In any event, we note that the modal intensity of consumption, conditional on non-zero, is 2 liters per household per week regardless of the socioeconomic group and the region. One can intuitively interpret this pervasive modal intensive margin as "one 2-liter family-size bottle of soda that is brought to the table every week."

Finally, we performed all manner of "consistency checks," where applicable, to ensure that the data were consistent across the different sources. For example, according to the HEX, among households residing in the three surveyed metropolitan areas in the Northeast, (statistically weighted) premium and generic market shares amount to 12.5% and 0.4%, respectively. (These shares are as defined in Section 2, including the outside option, and grow to 26.6% and 0.7%, respectively, if we condition on the 36% of households who have **ABC** socioeconomic status, comparable to the extensive margins reported in Table 2.) The (unconditional, three-city, 95/96) HEX shares of 12.5% and 0.4% in the example are similar to the Nielsen market shares of 12.3% across premium brands and 0.3% for generics in the Dec-96/Jan-97 bimonth (soda sold in family-size bottles through self-service outlets in the Northeast). By way of another example, the (projected) universe of households for region 3 (the metropolitan area of Rio de Janeiro) is 2.96 million under the HEX 95/96 (see Table 2), to be compared to 2.64 million households under the IBGE PNAD for Dec-96 (noting that Nielsen's region 3, which we adopt for the IBGE household counts, excludes some peripheral villages around the city of Rio de Janeiro). Further, using the HEX 95/96's balance sheet data, as explained in the text, we assigned **ABC** socioeconomic status to 55% of region 3's households (see Table 2), whereas the IBOPE suggest that at that time 57% of region 3's households were **ABC**.

A.2 Further details

A.2.1 Dynamic type evolution

We provide examples, from the data, of the dynamic updating process. We consider two transitions, both for region 4 (São Paulo Metro). The first transition, from $t = 1$ to $t = 2$, features upward mobility and, unusually in the data (yet we need to allow for this), a slight flow of residents out of the region ("net urban-to-rural migration"). The second transition, from $t = 10$ to $t = 11$, features upward mobility and the rural-to-urban migration that is prevalent in the data. We illustrate these transitions at the estimated model parameters θ^*. We also comment on the robustness of our estimates to the baseline mobility Assumptions 1 and 2.

Region 4, $t = 1$ to $t = 2$**.** The initial type-distribution vector is

$$\begin{aligned}\mathcal{F}_{g=4,t=1} &= \left\{F_{EA^{\mathcal{A}},4,1}, F_{EA^{\mathcal{B}},4,1}, F_{EA^{\mathcal{O}},4,1}, F_{NA^{\mathcal{A}},4,1}, F_{NA^{\mathcal{B}},4,1}, F_{NA^{\mathcal{O}},4,1}, F_{P^{\mathcal{A}},4,1}, F_{P^{\mathcal{B}},4,1}, F_{P^{\mathcal{O}},4,1}\right\} \\ &= \{.255, .029, .361, 0, 0, 0, .048, .009, .299\}\end{aligned}$$

As explained, the last element, say, is the product of (region 4's) poor household count in $t = 1$ (observed from IBOPE/IBGE) and the share of the outside option among region 4's **DE** households (calculated from the HEX 95/96), divided by the total household count (IBOPE/IBGE), i.e., $1346585 \times .84093/3789771 \simeq .299$. From $s_{j,r,g=4,t=1}(\theta^*)$ (see (4)), we obtain the mass of households for each of the nine types who choose to consume premium, generic, or no soda. For example, the share of premium soda among established affluent households who have a premium habit, $\sum_{j\in\mathcal{A}} s_{j,EA^{\mathcal{A}},g=4,t=1}(\theta^*) \simeq 97\%$. In contrast, the shares of premium soda among established affluents with generic habits and no-soda habit are 1% and 15%, respectively. Thus, since the established affluent population is constant over time (at 2443186), the number of established affluent households going into $t = 2$ with premium soda habits is (in thousands, hereafter) $3790\,(.255 \times .97 + .029 \times .01 + .361 \times .16) \simeq 1143$.

As for mobility, according to IBOPE/IBGE, the socioeconomic distribution of households evolves from (**ABC**,**DE**) $= (2443, 1347)$ in $t = 1$ to $(2511, 1269)$ in $t = 2$. It follows that, in $t = 2$: (i) $2511 - 2443 = 68$ households are newly affluent; (ii) $(2443 + 1347) - (2511 + 1269) = 10$ households migrated out of the urban area (again, this rarely happens in the data); and (iii) 1269 households are poor. Following Assumption 1 (Socioeconomic Mobility), the 68 upwardly mobile households entering $t = 2$ are endowed with habits in proportion to the choices of poor households in $t = 1$ among premium, generic and no soda (where these proportions are calculated as illustrated for established affluents, for which a proportion $1143/2443 \simeq 47\%$ chose premium rather than generic or no soda). These counts (summing 68) are deducted from the $t = 1$ poor population (1347) that is transitioning to $t = 2$ in proportion to the poor's choices across brand types. Similarly, following Assumption 2 (Migration), the 10 households leaving the city are

dropped from the counts of the poor (totaling $1347 - 68$) in proportion to the poor's choices across brand types.

Region 4, $t = 10$ to $t = 11$. We keep this example brief, highlighting mobility. The type-distribution vector following choices made in $t = 9$ and mobility into $t = 10$ is

$$\mathcal{F}_{g=4,t=10} = \{.233, .099, .306, .019, .016, .110, .000, .002, .213\}$$

Having updated from $t = 1$, the history of choices and mobility now determines the distribution of habits across each socioeconomic group. By $t = 10$, the "premium-to-generic ratio" is $.019 : .016 = 1.2$ among newly affluent households, compared to $.233 : .099 = 2.4$ among the established affluent (see Table 5). From IBOPE/IBGE data, the mass of households by socioeconomic group (in thousands) in $t = 10$ is computed as: 2443 established affluent (this stays constant), 560 newly affluent and 826 poor (see Figure 4; $t = 10$ is the Jun-98/Jul-98 bimonth).

The evolution of (**ABC**,**DE**) from $(3003, 826)$ in $t = 10$ to $(3060, 784)$ in $t = 11$ implies that: (i) the newly affluent count grows by 57 (to 617); (ii) 16 migrants arrive at the city and join the ranks of the poor; and (iii) the poor count drops by $57 - 16 = 42$ (to 784). The 57 upwardly mobile households making choices with newly affluent status in $t = 11$ are endowed with habits in proportion to the $t = 10$ choices of the poor they left behind (Assumption 1). The 16 migrants who are new to the city have a no-soda habit (Assumption 2).

Robustness to Assumptions 1 and 2. Our results are robust to alternative mobility assumptions, namely: (i) modifying Assumption 1 to endow households moving up from poor to newly affluent status with habits in proportion to the previous-period soda choices of the newly affluents they are joining, rather than the poor they are leaving behind[33] (and analogously with respect to households moving down from newly affluent to poor status, based on the previous-period choices of the poor); and (ii) modifying Assumption 2 to endow households moving to urban areas with habits in proportion to the previous-period soda choices of the city-dwelling poor they are joining. For example, under (ii), $(\alpha_{EA}, \alpha_{NA}, \alpha)$ and λ are estimated, respectively, at $(2.98, 2.77, -5.27)$ and 5.09 (with standard errors of $(1.43, 1.23, 1.45)$ and .37), very close to baseline estimates (see Table 3). Full estimates of these model variants are available upon request.

A.2.2 The estimation algorithm

In what follows, we explain the structure of the GMM objective, and then detail how this objective function is evaluated at some generic value for the parameters $\theta = (\theta_1, \theta_2)$.

Given any generic value for the non-linear parameters θ_2, steps 1 to 5 of the algorithm below

[33] The exception is the first transition, from $t = 1$ to $t = 2$, in which the newly affluent are a random sample of the poor as, by definition, there are no newly affluents in $t = 1$.

yield an $N \times 1$ vector $\delta(\theta_2)$, containing the base-utility levels for all brands in all regions in all time periods ($N = 9 \cdot 7 \cdot 57$). As noted in Section 4, conditioning on the full parameter vector θ, one obtains an $N \times 1$ vector of base-utility unobservables by subtracting the systematic portion of the base utility from δ_{jgt}, i.e., $\xi_{jgt} = \delta_{jgt} - x'_{jgt}\beta - \alpha \cdot p_{jgt}$. Stacking all these unobservables together, we can write:

$$\xi(\theta) = \delta(\theta_2) - X\theta_1$$

where the $N \times K_1$ matrix X contains the K_1 base-utility covariates (including price), and let K_2 denote the dimension of θ_2. Now let Z denote a $N \times L$ matrix of instruments containing all covariates in X but price, as well as excluded instruments (e.g., cost shifters), where $L > K_1 + K_2$. Writing $W = (Z'Z)^{-1}$, the GMM objective is defined by:

$$Q_N(\theta) = \xi(\theta)' Z W Z' \xi(\theta)$$

Computation time can be reduced substantially by noting (see BLP 1995, Nevo 2000) that, conditional on a guess for θ_2, there is a closed-form solution for the parameters θ_1 that minimizes the objective:

$$\theta_1^*(\theta_2) = \left(X'ZWZ'X\right)^{-1} X'ZWZ'\delta(\theta_2)$$

This allows us to maximize the objective by searching only over values of the non-linear parameters θ_2.

At every guess $\tilde{\theta}_2$ for the non-linear parameters, the GMM objective is evaluated via the following steps:

1. For every region $g = 1, ..., 7$, and period $t = 1$, given $\tilde{\theta}_2$ and $\mathcal{F}_{g1}$, use the BLP contraction mapping to solve for the unique vector of base utilities that matches observed aggregate market shares with those predicted by the model.

2. For every region $g = 1, ..., 7$ and household type $r = 1, ..., 9$, use equation (4), the base utilities recovered in step 1, and $\tilde{\theta}_2$, to predict the shares of type-r households who consume premium brands, generic brands or no soda in period $t = 1$.

3. For every region $g = 1, ..., 7$, use the shares obtained in step 2, data on aggregate social mobility, and Assumptions 1 and 2, to forward-update the proportion of households in period $t = 2$ who belong to each of the nine types, $\mathcal{F}_{g2}$ (see Section 3.2).

4. Repeat steps 1-3 for periods $t = 2, ..., 57$.

5. Stack the base-utility vectors for all brands, time periods and regions in the $N \times 1$ vector $\delta\left(\tilde{\theta}_2\right)$, and evaluate the GMM objective at the guess $\tilde{\theta}_2$, as explained above.

A.2.3 Robustness

Given space restrictions, we briefly describe some of the alternative specifications, on top of the alternative mobility assumptions discussed above, that we have estimated to confirm the validity and robustness of our baseline results. Estimates of these model variants are available upon request.

Market size. Our baseline model defines market potential as six liters per week, interpreted as 3 meals/week in which a 2-liter family-size bottle of soda might be brought to the table. Estimated price sensitivities and the habit parameter hardly vary as we vary the number of meals per week between 2.6, 2.7, ... , 3.3. Beyond this range, estimated $(\alpha_{EA}, \alpha_{NA}, \alpha)$ vary more, but our estimate for λ is very stable about 5 (all the way from 2.0 to 3.6 meals/week).

Habit formation. Specifications that we implemented, each addressing alternative mechanisms than the one we wish to highlight, include: (i) allowing habit to form for soda in general, *regardless* of the type of brand (i.e., consuming either heavily advertised premium or discount generic soda in this period shifts the utility from consuming any soda next period by λ); (ii) allowing loyalties to form for the flagship premium brands Coke (including Diet Coke), Guaraná Antarctica, Fanta, or Pepsi (i.e., consuming Pepsi in this period increases one's utility from consuming Pepsi in the next period—but not another brand—by λ); and (iii) allowing loyalty to form only for the Coke (including Diet Coke) brand. To illustrate, model (ii) has five habit states which, interacted with 3 socioeconomic groups, implies 15 household types (and we must modify the initial conditions from the HEX accordingly). Brand loyalty is estimated to be strong and significant under alternative models (ii) and (iii), leading to aggregate own-price elasticities that appear too low in magnitude, namely, -0.7 and -0.9 for Coke under (ii) and (iii), respectively (compared to -1.7 in Table 4). In general, estimated habit parameter(s) under these alternative models are large and significant, but do not provide as strong a justification for Coca-Cola's mid 1999 price cut.

Other specifications. A more general model allowed the premium habit and the frugal habit to vary in magnitude. Estimated habit parameters $\lambda^{\mathcal{A}}$ and $\lambda^{\mathcal{B}}$, for premium and frugal respectively, are 5.31 (s.e. 0.41) and 4.65 (s.e. 0.50). We also tested robustness with regard to: (i) initial HEX 95/96 shares (namely, expanding the HEX outlet codes that map to Nielsen's stores with checkouts); and (ii) defining market share by the extensive margin once the region-specific intensive margin, as observed in the single cross-section of household-level data (HEX 95/96), is fixed over time.

A.2.4 Variable profit

Our back-of-the-envelope calculation of variable profit considers the three-year period between April 2000 and March 2003. We assume that the premium sellers' net sales price is 35% of the price Nielsen observes on the shelf, which is paid by the end consumer. (See Ambev 2003 and Salvo 2009 for a discussion of the very high taxes incurred along the formal vertical chain, as well as vertical relations. We also base our calculations on interviews with an executive at a premium seller.) Thus, the observed shelf price of R$ 0.913 / liter (sales weighted across premium brands, averaged over the three years) corresponds to a net sales price for Coca-Cola/Ambev of R$ 0.320 / liter, net of sales tax, retail margin, and distribution costs. Had the premium sellers not cut prices in mid 1999, we assume that this price would have been proportionately higher, at R$ 0.384 / liter. Based on Ambev (2003), we take the "cost of goods sold" as R$ 0.199 / liter. We note that the real prices of sugar, plastic, electricity, labor, and fuel were quite stable between 2000 and 2002 (in general, they began rising at the end of our sample period, in 2003). The variable profit margins for the Coca-Cola/Ambev "systems" are thus R$ 0.120 / liter with the observed price cut and R$ 0.185 / liter under the counterfactual of no price cut. Multiplying by the premium sellers' observed and counterfactual quantities sold over this three-year period (namely, 7.2 billion liters observed; 4.0 bi liters counterfactual under the Brand Type Persistence model; 5.4 bi liters counterfactual under the "no habit" model variant) yields the variable profits stated in the text (respectively, R$ 860 million, R$ 740 million, and R$ 1.0 billion).

B Tables and Figures

Table 1: Brand Volume Shares of the Soda Category

Brand	Region 1		Region 2		Region 3		Region 4		Region 5		Region 6		Region 7	
	t=1	t=57	t=1	t=57	t=1	t=57	t=1	t=57	t=1	t=57	t=1	t=57	t=1	t=57
Coke	0.40	0.21	0.37	0.26	0.36	0.26	0.32	0.24	0.31	0.25	0.28	0.31	0.34	0.29
Fanta	0.08	0.06	0.05	0.08	0.05	0.09	0.08	0.09	0.06	0.08	0.06	0.08	0.03	0.05
Kuat	0.03	0.06	0.02	0.04	0.04	0.06	0.07	0.06	0.06	0.04	0.04	0.04	0.05	0.03
Diet Coke	0.02	0.03	0.02	0.03	0.03	0.05	0.02	0.04	0.03	0.03	0.03	0.04	0.02	0.03
Other Coca-Cola	0.02	0.03	0.02	0.02	0.02	0.02	0.05	0.06	0.04	0.05	0.05	0.04	0.05	0.02
Guarana Antartica	0.17	0.09	0.07	0.06	0.06	0.09	0.09	0.10	0.08	0.07	0.05	0.05	0.09	0.10
Other Ambev	0.19	0.02	0.11	0.01	0.15	0.01	0.16	0.02	0.12	0.01	0.17	0.01	0.06	0.02
Pepsi	0.06	0.03	0.08	0.03	0.19	0.07	0.05	0.05	0.06	0.04	0.11	0.09	0.09	0.05
Generics	0.03	0.47	0.26	0.47	0.12	0.36	0.17	0.34	0.25	0.41	0.21	0.33	0.27	0.41

Volume shares of the soda category, by brand in each region in the first and last time periods. Coke, Fanta, Kuat, Diet Coke, and "Other Coca-Cola" are premium brands marketed by the Coca-Cola Company. Guarana Antarctica, Pepsi, and "Other Ambev" are premium brands marketed by Ambev. Source: Nielsen.

Table 2: Soda Consumption by Socioeconomic Group (HEX)

Region of survey. cities	Socioeconomic group	Households ×1000 Universe	%	Soda purchasing	By brand type Premium	Generic	No soda
1 (Northeast)	ABC	696	36	28.0%	27.0%	0.9%	72.0%
	DE	1230	64	9.1%	8.3%	0.8%	90.9%
2 (MG, ES,	ABC	529	57	39.9%	37.9%	2.0%	60.1%
RJ interior)	DE	404	43	23.2%	22.1%	1.2%	76.8%
3 (RJ Metro)	ABC	1625	55	31.9%	31.6%	0.3%	68.1%
	DE	1331	45	18.3%	18.3%	0.0%	81.7%
4 (SP Metro)	ABC	2586	60	34.5%	33.1%	1.4%	65.5%
	DE	1689	40	19.8%	17.3%	2.6%	80.2%
6 (South)	ABC	955	63	43.2%	42.5%	0.7%	56.8%
	DE	559	37	20.4%	20.1%	0.3%	79.6%
7 (DF, GO	ABC	428	61	36.5%	34.4%	2.1%	63.5%
MS)	DE	270	39	23.6%	21.1%	2.5%	76.4%
Total above	ABC	6819	55	35.0%	33.9%	1.1%	65.0%
	DE	5482	45	17.5%	16.3%	1.2%	82.5%

The extensive margin of soda consumption inside the home by different socioeconomic groups in 1995/96. Socioeconomic groups are defined per the points scale used by IBOPE. Metropolitan areas surveyed were: (Region 1) Recife, Fortaleza and Salvador; (Region 2) Belo Horizonte; (Region 3) Rio de Janeiro Metro; (Region 4) Sao Paulo Metro; (Region 6) Curitiba and Porto Alegre; (Region 7) Brasilia and Goiania. No city was surveyed in Region 5 (state of Sao Paulo excluding Sao Paulo Metro). We do not consider the northern city of Belem as it is located outside the area covered by Nielsen. Source: IBGE HEX (POF) 1995/96.

Table 3: Demand Estimation Results

Price Sensitivity Parameters					
α_{EA}	2.96	(1.42)			
α_{NA}	2.75	(1.22)			
α	-5.25	(1.44)			
Habit Parameter					
λ	5.09	(0.37)			
Additional Covariates					
Constant	-3.24	0.30			
Temperature	3.06	(0.31)			
Advertising Effects:			Region-specific Time Trends:		
Advertising GRPs×Region 1	1.06	(0.46)	Region 1	-0.45	(0.18)
Advertising GRPs×Region 2	0.83	(0.63)	Region 2	-0.21	(0.20)
Advertising GRPs×Region 3	0.26	(0.29)	Region 3	-0.61	(0.16)
Advertising GRPs×Region 4	0.41	(0.45)	Region 4	-0.88	(0.12)
Advertising GRPs×Region 5	0.62	(0.45)	Region 5	-0.42	(0.16)
Advertising GRPs×Region 6	0.72	(0.40)	Region 6	-0.30	(0.15)
Advertising GRPs×Region 7	0.50	(0.35)	Region 7	-0.52	(0.16)
Seasonality× Brand Type Effects	Yes				
Brand-Region Fixed Effects	Yes				

Standard errors in parentheses. Source: estimated baseline model.

Table 4: Estimated Demand Elasticities

Aggregate Own-Price Elasticities		Household-Type Specific Elasticities			
Coke	-1.69	Coke, $EA^{\mathcal{A}}$	-1.51	Generic, $EA^{\mathcal{A}}$	-1.43
Guaraná Antarctica	-2.13	Coke, $EA^{\mathcal{B}}$	-2.46	Generic, $EA^{\mathcal{B}}$	-0.19
Fanta	-2.07	Coke, $EA^{\mathcal{O}}$	-2.39	Generic, $EA^{\mathcal{O}}$	-1.37
Pepsi	-2.07	Coke, $NA^{\mathcal{A}}$	-1.70	Generic, $NA^{\mathcal{A}}$	-1.56
Generic	-0.51	Coke, $NA^{\mathcal{B}}$	-2.69	Generic, $NA^{\mathcal{B}}$	-0.23
		Coke, $NA^{\mathcal{O}}$	-2.63	Generic, $NA^{\mathcal{O}}$	-1.50
		Coke, $P^{\mathcal{A}}$	-4.98	Generic, $P^{\mathcal{A}}$	-3.29
		Coke, $P^{\mathcal{B}}$	-5.66	Generic, $P^{\mathcal{B}}$	-1.52
		Coke, $P^{\mathcal{O}}$	-5.65	Generic, $P^{\mathcal{O}}$	-3.27

Reported elasticities are means across region-and-time markets, as predicted by the estimated demand model. Only a few elasticities are shown.

Table 5: Predicted Soda Penetration and Consumption Patterns by Socioeconomic Group

	Soda Penetration	Premium Share	Generic Share	Premium:Generic Ratio
Established Affluent	0.51	0.34	0.17	2.0
Newly Affluent	0.37	0.22	0.15	1.5
Poor	0.03	0.01	0.02	0.6

Soda penetration, for the entire category and by type of brand, in each socioeconomic group, as predicted by the estimated demand model. Reported numbers are means across region-and-time markets.

Table 6: Estimates from the "No Habit" Demand Model

	coeff	(s.e.)
Price Sensitivity Parameters		
α_{EA}	1.90	(0.70)
α_{NA}	-0.19	(0.55)
α	-3.34	(0.73)
Additional Covariates		
Constant	-1.25	(0.14)
Temperature	1.76	(0.16)
Advertising× Region	Yes	
Region-specific time trends	Yes	
Seasonality× Brand Type Effects	Yes	
Brand-Region Fixed Effects	Yes	

Source: the no-habit model variant.

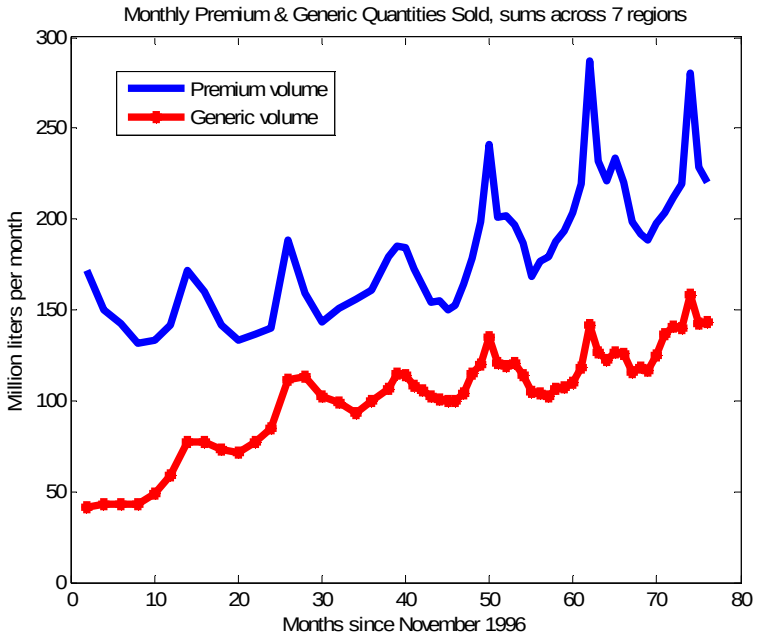


Figure 1: The evolution of quantities (in million liters / month) by type of brand (Premium versus Generic), for soda sold in family-size bottles through the self-service channel across the seven Nielsen regions, in the period Dec-96 to Mar-03. Source: Nielsen.

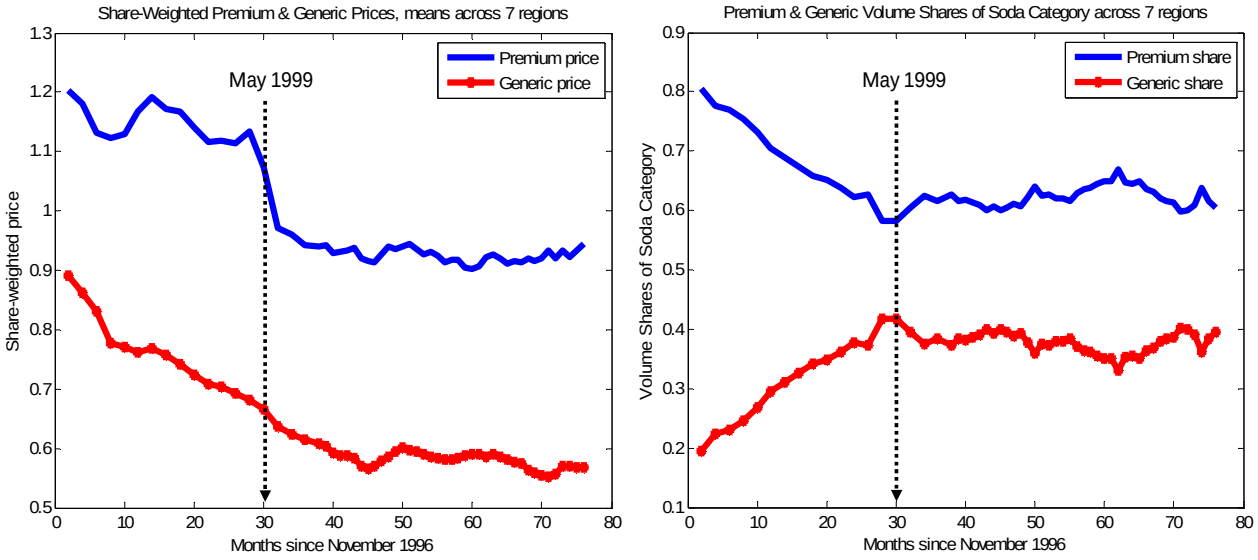


Figure 2: The evolution of prices (in constant Brazilian R$ / liter) and category volume shares (in percent, summing to one) by type of brand (Premium versus Generic), for soda sold in family-size bottles through the self-service channel across the seven Nielsen regions, in the period Dec-96 to Mar-03. Source: Nielsen.

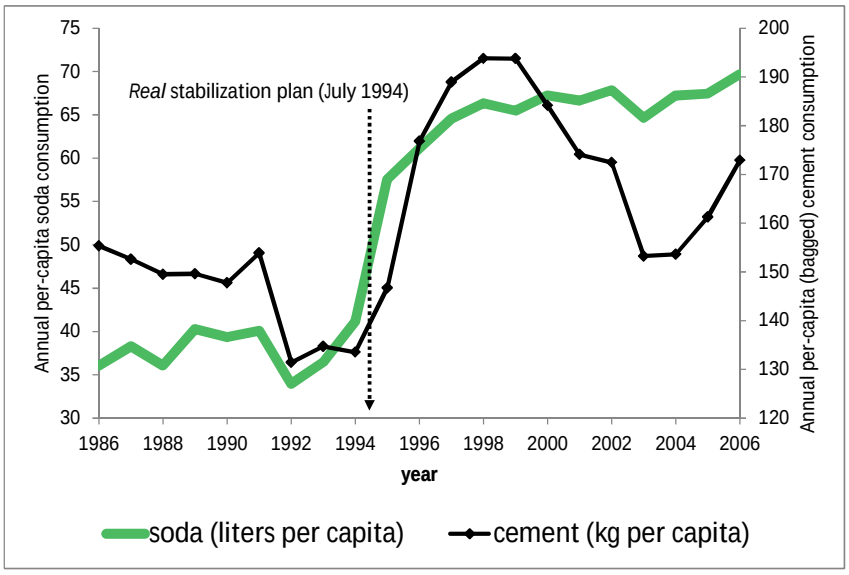


Figure 3: Annual aggregate per capita consumption of soft drinks (in liters per person) and of bagged cement (in kilograms per person). (Cement sold in bags, as opposed to sales in bulk, filter out any large-scale construction activity, such as government spending on infrastructure.) Source: Brazilian trade associations for soft drink makers and for cement producers, ABIR and SNIC respectively.

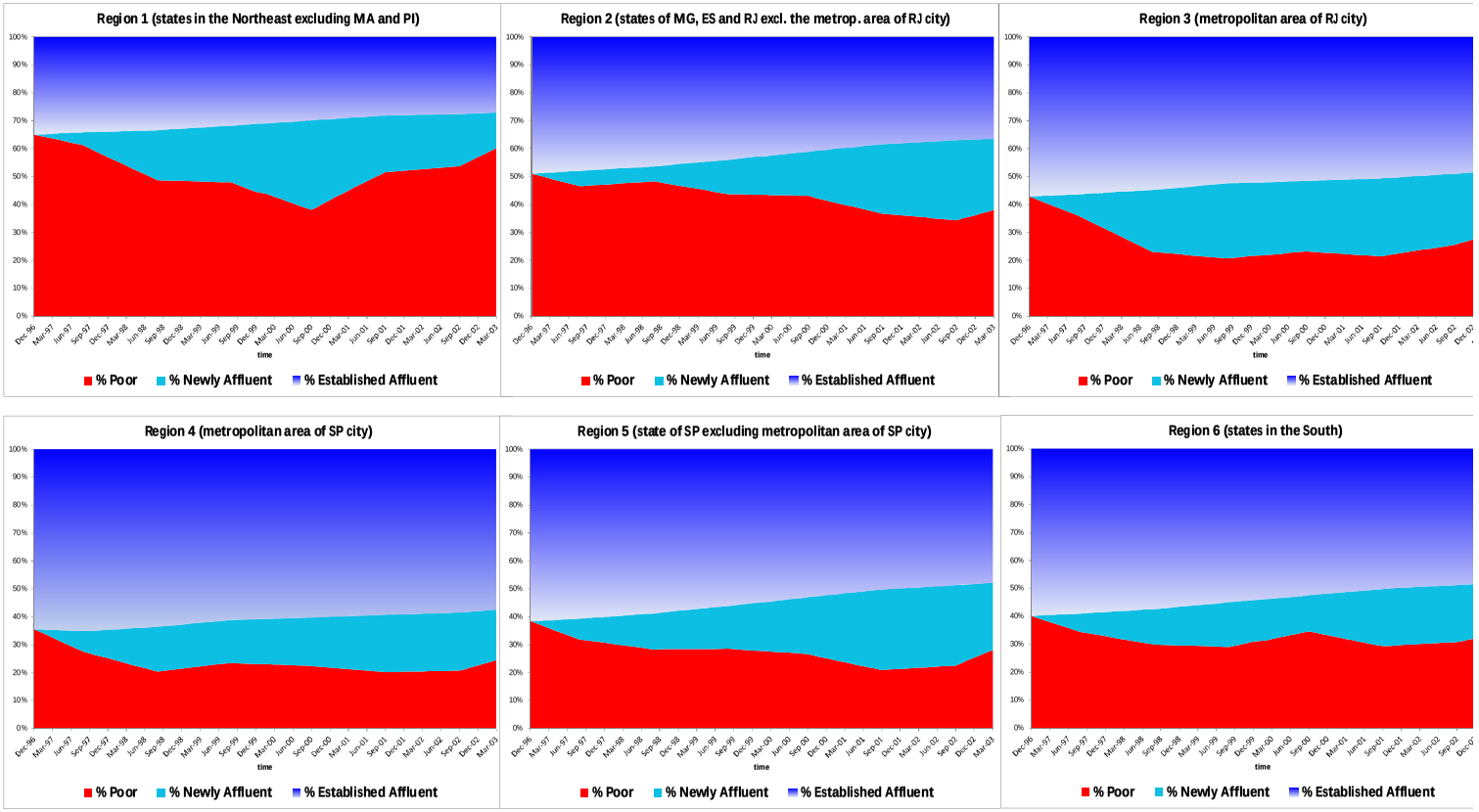


Figure 4: The rise of "newly affluent" households: Proportion of urban households in each of three constructed socioeconomic groups ("Established Affluent," "Newly Affluent" and "Poor"), as defined in the text, by region in the period Dec-96 to Mar-03. The smallest region by number of households, region 7 (Federal District and states of GO and MS), is not shown for lack of space (the pattern is similar to region 2). Source: IBOPE LatinPanel and IBGE PNAD.

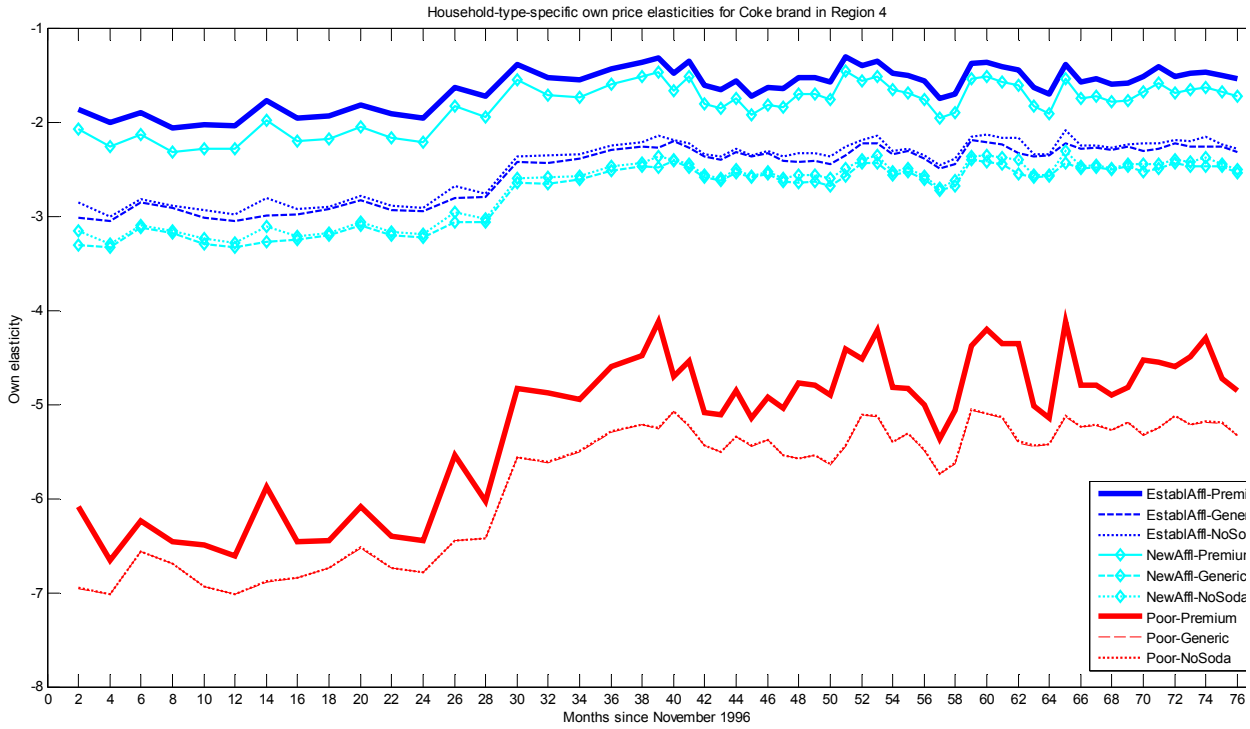


Figure 5: Evolution of own-price elasticities for Coke brand, by household type, in region 4 (São Paulo Metro). Source: Baseline model (Brand Type Persistence). Source: baseline model.

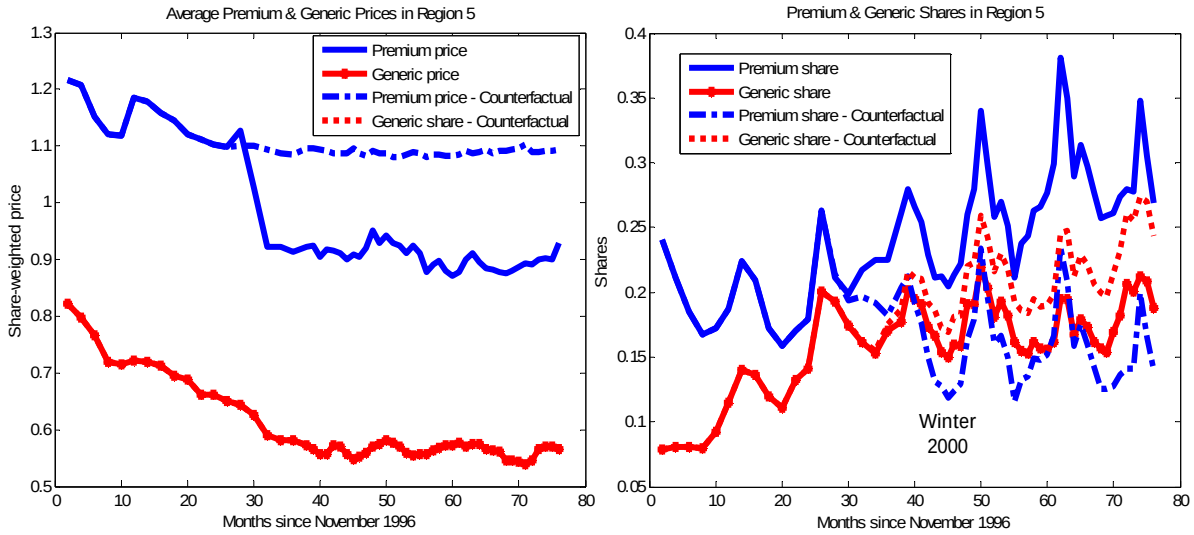


Figure 6: Actual against counterfactual price and share paths for premium brands and generic brands in region 5 (São Paulo Interior). Prices in the left panel and shares in the right panel. The counterfactual scenario considers premium brands not cutting prices in mid 1999. Source: Baseline model (Brand Type Persistence).

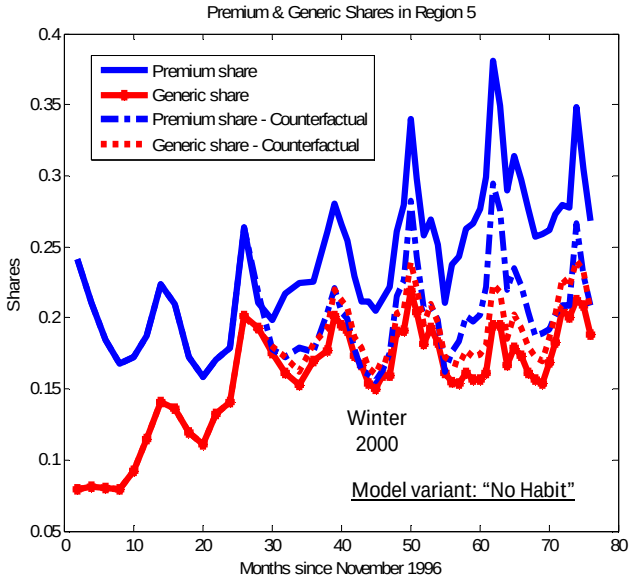


Figure 7: Share paths for premium brands and generic brands, in region 5, for the same counterfactual experiment of the earlier figure (premium brands not cutting prices in 1999), but employing the "No Habit" model variant.

Borders, Geography, and Oligopoly: Evidence from the Wind Turbine Industry*

A. Kerem Cosar
Chicago Booth

Paul L. E. Grieco
Penn State

Felix Tintelnot
Penn State

April 5, 2012

Abstract

Using a micro-level dataset of wind turbine installations in Denmark and Germany, we estimate a structural oligopoly model with cross-border trade and heterogeneous firms. Our approach separately identifies border-related from distance-related variable costs and bounds the fixed cost of exporting for each firm. Variable border costs are large: equivalent to roughly 400 kilometers (250 miles) in distance costs, which represents 40 percent of the average exporter's total delivery costs. Fixed costs are also important; removing them would increase German firms' market share in Denmark by 10 percentage points. Counterfactual analysis indicates that completely eliminating border frictions would increase total welfare in the wind turbine industry by 5 percent in Denmark and 10 percent in Germany.

JEL Codes: F14, L11, L20, L60, R12
Keywords: trade costs, oligopoly, spatial competition, constrained MLE

*For their insightful comments, we thank Costas Arkolakis, Lorenzo Caliendo, Allan Collard-Wexler, Steve Davis, Jonathan Eaton, Charles Engel, Chang-Tai Hsieh, Oleg Itskhoki, Samuel Kortum, Kala Krishna, Matthias Lux, Ferdinando Monte, Eduardo Morales, Brent Neiman, Dennis Novy, Ralph Ossa, Ayşe Pehlivan, Joris Pinkse, Mark Roberts, Andrés Rodríguez-Clare, Alan Spearot, Elie Tamer, Jonathan Vogel, Stephen Yeaple, Nicholas Ziebarth, and workshop participants at Penn State, Chicago Booth, Boston University, Boston FED, UCLA, U.S. Census Bureau, Ohio State, University of Minnesota, University of Munich, Koç University, Massachusetts Institute of Technology, Columbia University, Princeton University, EIIT in Purdue University, Midwest Trade Conference, LACEA-TIGN, Conference on Globalization in Denmark, and UECE Lisbon Meetings in Game Theory and Applications. For their valuable discussions with us about the industry, we thank Wolfgang Conrad, Robert Juettner, Alex Kaulen, and Jochen Twele.

Correspondence: kerem.cosar@chicagobooth.edu, paul.grieco@psu.edu, fut111@psu.edu.

1 Introduction

Distance and border barriers lead to geographical and national segmentation of markets. In turn, the size and structure of markets depend crucially on the size and nature of trade costs. A clear understanding of these costs is thus important for assessing the impact of many government policies.[1] Since the seminal work of McCallum (1995), an extensive literature has documented significant costs related to crossing national boundaries. Estimated magnitudes of border frictions are so large that some researchers have suggested they are due to spatial and industry-level aggregation bias, a failure to account for within-country heterogeneity and geography, and cross-border differences in market structure.[2] To avoid these potentially confounding effects, we use spatial micro-data from wind turbine installations in Denmark and Germany to estimate a structural model of oligopolistic competition with border frictions. Our main findings are: (1) border frictions are large within the wind turbine industry, (2) fixed and variable costs of exporting are both important in explaining overall border frictions, and (3) these frictions have a substantial impact on welfare.

Our ability to infer various components of trade costs is a result of our focus on a narrowly defined industry: wind turbine manufacturing. In addition to being an interesting case for study in its own right due to the growing importance of wind energy to Europe's overall energy portfolio, the wind turbine industry in the European Union (EU) offers an opportunity to examine the effects of national boundaries on market segmentation. First, we have rich spatial information on the location of manufacturers and installations. The data are much finer than previously used aggregate state- or province-level data. The use of disaggregated data allows us to account for actual shipping distances, rather than rely on market-to-market distances, to estimate border costs. Second, the data contain observations of both domestic and international trade. We observe active manufacturers on either side of the Danish-German border, some of whom choose to export and some of whom do not, allowing us to separate fixed and variable border costs. Third, intra-EU trade is free from formal barriers and large exchange rate fluctuations. It is also subject to wide-ranging efforts to minimize informal barriers.[3] By the Single European Act, national subsidies are directed only toward the generation of renewable electricity and do not discriminate against other European producers of turbines. The border costs in this setting are therefore due to factors other than formal barriers to trade and exchange rate fluctuations.

[1] Policy relevance goes beyond trade policies. According to Obstfeld and Rogoff (2001), core empirical puzzles in international macroeconomics can be explained as a result of costs in the trade of goods. Romalis (2007) shows that the interaction of tax policies and falling trade costs was key to the rapid growth of Ireland in the 1990s. Effectiveness of domestic regulation in some industries may hinge on the extent of trade exposure, as shown by Fowlie, Reguant, and Ryan (2011) for the US Portland cement industry.

[2] See Hillberry (2002), Hillberry and Hummels (2008), Broda and Weinstein (2008) and Gorodnichenko and Tesar (2009).

[3] All tariffs and quotas between former European Economic Community members were eliminated by 1968. The Single European Act came into force in 1987 with the objective of abolishing all remaining physical, technical and tax-related barriers to free movement of goods, services, and capital within the EU until 1992. Between 1986 and 1992, the EU adopted 280 pieces of legislation to achieve that goal.

Despite major formal integration, the data indicate substantial market segmentation between Denmark and Germany. Examining the sales of turbines in 1995 and 1996, we find that domestic manufacturers had a substantially higher market share than did foreign manufacturers. For example, the top five German manufacturers possessed a market share of 60 percent in Germany and only 2 percent in Denmark. The market share of Danish producers drops by approximately 30 percent at the border.

What are the sources of cross-national market segmentation? On one hand, a cursory glance at our data suggests that national borders affect the decisions of firms to enter the foreign market. To be specific, only one of the five large German firms exports to Denmark. On the other hand, all five large Danish firms have sales in Germany, but their market share is substantially lower in the foreign market and drops discontinuously at the border. The difference in participation patterns across the border reflects fixed costs faced by exporting firms. The change in market share at the border may be generated by differences in competition (e.g., differences in the set of competitors and their underlying characteristics) or by higher variable costs for foreign firms. To explain differences in market shares along extensive and intensive margins, we propose a model of cross-border oligopolistic competition that embeds costs for exporting as primitives and controls for other sources of cross-border differences. This allows us to infer the costs that exporter firms face and quantify their impact on market shares, profits, and consumer welfare through counterfactual analysis.

In our model, firms are heterogeneous in their production costs, foreign market entry costs, and distance to project sites. To become active in the foreign country, firms must pay a *fixed border cost* specific to them. Fixed border costs include maintaining a foreign sales force, developing technology to connect turbines to the foreign electricity grid, and gaining certification for turbine models in foreign countries. The model incorporates two types of variable costs for supplying a project: First, all firms face a *distance cost* that increases with the distance between the location where they produce the turbines and the location of the project. Second, exporters pay an additional *variable border cost* to supply projects in their foreign market. While the distance friction is analogous to the standard iceberg cost in trade models, the variable border cost captures additional hurdles that exporters face independent of shipping distance. These hurdles may arise due to language or cultural differences between purchasers and manufacturers, legal complications due to the use of cross-border contracts, or the need to interact with multiple national transportation authorities to authorize turbine delivery. One of our objectives is to gauge the importance of each type of cost in segmenting the markets.

The model has two stages: In the first stage, turbine producers decide whether or not to export. This depends on whether their expected profit in the foreign market exceeds their fixed border cost. As a result, the set of competing firms changes at the border. In the second stage, turbine producers observe the set of

active producers in each market and engage in price competition for each project. A producer's costs depend on the location of the project through both distance and the presence of a border between the producer and the project. For each project, firms choose prices (and hence, markups) on the basis a profit maximization condition derived from our model. Project managers then face a discrete choice problem: they observe price bids and pick the producer that maximizes their project's value. In equilibrium, each firm takes into account the characteristics of its competitors when choosing its own price. The model thus delivers endogenous variation in prices, markups, and market shares across points in space. Our data informs us about the suppliers of all projects. We estimate the model by maximizing the likelihood of correctly predicting these outcomes.

Our results indicate that there are substantial costs to sell wind turbines across the border between Denmark and Germany. We find that the variable border costs are roughly equivalent to moving a manufacturer 400 kilometers (250 miles) further away from a project site. Given that the largest possible distance from the northern tip of Denmark to the southern border of Germany is roughly 1,400 kilometers (870 miles), this is a significant cost for foreign firms. Removing fixed costs of foreign entry, such that all firms compete on both sides of the border, raises the market share of German firms in Denmark from 2 to 12 percent; also eliminating variable border costs raises that market share from 12 percent to 22 percent. Counterfactual analysis provides further insights into the welfare effects of borders. A hypothetical elimination of all border frictions raises consumer surplus by 10.4 and 15.3 percent in Denmark and Germany, respectively. Removing border frictions increases profits of foreign firms while reducing those of domestic firms. The net effect is small in Denmark (producer surplus declines by less than 1 percent) and large in Germany (producer surplus declines by over 6 percent). Overall, consumer gains outweigh producer losses in both countries. Total surplus increases by 5 percent in Denmark and 10 percent in Germany.

The paper adds to the empirical literature on trade costs by estimating a structural oligopoly model that controls for internal geography and firm heterogeneity. McCallum (1995) and Anderson and van Wincoop (2003) use data on interstate, interprovincial, and international trade between Canada and the United States to document a disproportionately high level of *intranational* trade between Canadian provinces and U.S. states after controlling for income levels of regions and the distances between them. Engel and Rogers (1996) find a high level of market segmentation between Canada and the United States using price data on consumer goods. Gopinath, Gourinchas, Hsieh, and Li (2011) use data on retail prices to document large retail price gaps at the border using a regression discontinuity approach. Goldberg and Verboven (2001, 2005) find considerable price dispersion in the European car market and some evidence that the markets are

becoming more integrated over time.[4]

Rather than inferring a "border effect" or "width of the border" based on differences between intra- and international trade flows and price differentials, we use spatial micro-data to estimate trade costs which induce market segmentation. By doing this, we addresses several critiques raised by the literature. Hillberry (2002) and Hillberry and Hummels (2008) show that sectoral and geographical aggregation lead to upward bias in the estimation of the border effect in studies that use trade flows. Holmes and Stevens (2012) emphasize the importance of controlling for internal distances. In a similar fashion, Broda and Weinstein (2008) find that aggregation of individual goods' prices amplifies measured impact of borders on prices. Our data enables us to calculate the distances between consumption and production locations for a narrowly defined product. That, in turn, enables us to separate the impact of distance from the impact of the border.

Our structural model of oligopolistic competition controls for differences in market structure and competitor costs across space. The estimates from our structural model can thus be directly interpreted as costs that exporters must pay to market their products abroad.[5] This approach addresses the concern of Gorodnichenko and Tesar (2009) that model-free, reduced-form estimates fail to identify the border effect. To highlight the importance of using disaggregated data and a structural model, Section 6 presents an experiment based on our estimated model in which we regress imputed price differentials on distances and a border dummy to calculate the implied width of the border. This width is substantially larger than what our structural model implies. A comparison of structural and reduced form equations illustrates the sources of bias.

In summary, our industry-specific focus has three major advantages: First, the use of precise data on locations in a structural model allows for a clean identification of costs related to distance and border. Second, the model controls for endogenous variation in markups across markets within and across countries based on changes in the competitive structure across space. Third, by distinguishing between fixed and variable border costs, we gain a deeper insight into the sources of border frictions than we do from studies that use aggregate data.

In the following section, we discuss our data and provide background information for the Danish-German wind turbine industry. We also present some preliminary analysis that is indicative of a border effect. Section 3 introduces our model of the industry. We show how to estimate the model using maximum

[4]The interest in border frictions partially stemmed from the realization that prices of tradable goods do not immediately respond to exchange rate fluctuations, leading to substantial price differentials across countries. The exchange rate between Germany and Denmark was extremely stable during our sample period: the median month-to-month variation is 0.23 percent. So, this source of border frictions is absent from our environment.

[5]It may also be that preferences change at the border such that consumers act on a home bias for domestic turbines. In our setting, demand comes from profit maximizing energy producers buying an investment good, so we expect that demand driven home bias are less likely to occur than they would for a consumption good. Within our model, home bias in consumer preferences cannot be separately identified from border costs. Alternatively, we can interpret our results as incorporating the additional costs exporting firms must incur to overcome any home-bias in preferences.

likelihood with equilibrium constraints and present the results in Section 4. In Section 5, we perform a counterfactual analysis of market shares and welfare by re-solving the model without fixed and variable border costs. Section 6 uses market-to-market price differentials from our model in a reduced-form regression to relate our approach to studies that estimate border frictions based on the law of one price. We conclude in Section 7 with a discussion of policy implications.

2 Industry Background and Data

Encouraged by generous subsidies for wind energy, Germany and Denmark have been at the forefront of what has become a worldwide boom in the construction of wind turbines. Owners of wind farms are paid for the electricity they produce and provide to the electric grid. In both countries, national governments regulate the unit price paid by grid operators to site owners. These "feed-in-tariffs" are substantially higher than the market rate for other electricity sources. Important for our study is that public financial support for this industry is not conditional on purchasing turbines from domestic turbine manufacturers, which would be in violation of European single market policy. So, it is in the best interest of the wind farm owner to purchase the turbine that maximizes his or her profits independent of the nationality of the manufacturer.

The project owner's choice of manufacturer is our primary focus. In the period we study, purchasers of wind turbines were primarily independent producers, most often farmers or other small investors.[6] The turbine manufacturing industry, on the other hand, is dominated by a small number of manufacturing firms that both manufacture turbines and construct them on the project owner's land. Manufacturers usually have a portfolio of turbines available with various generating capacities. Overall, their portfolios are relatively homogeneous in terms of observable characteristics.[7] There could be, however, differences in quality and reliability that we do not directly observe.

The proximity of the production location to the project site is an important driver of cost differences. Due to the size and weight of turbine components, oversized cargo shipments typically necessitate road closures along the delivery route (see Figure 1). Transportation costs range between 6 to 20 percent of total costs (Franken and Weber, 2008). In addition, manufacturers usually include maintenance contracts as part of the turbine sale, so they must regularly revisit turbine sites after construction.

[6] Small purchasers were encouraged by the financial incentive scheme that gave larger remuneration to small, independent producers such as cooperative investment groups, farmers, and private owners. The German Electricity Feed Law of 1991 explicitly ruled out price support for installations in which the Federal Republic of Germany, a federal state, a public electricity utility or one of its subsidiaries held shares of more than 25 percent. The Danish support scheme provided an about 30% higher financial compensation for independent producers of renewable electricity (Sijm (2002)). A new law passed in Germany in 2000 eliminated the restrictions for public electricity companies to benefit from above market price renumeration of renewable energy.

[7] Main observable product characteristics are generation capacity, tower height, and rotor diameter. Distribution of turbines in terms of these variables is very similar in both countries. Further details are displayed in Table 8 in Appendix A.

Figure 1: Transportation of Wind Turbine Blades

Notes: A convoy of wind turbine blades passing through the village of Edenfield, England. Photo Credit: Anderson (2007)

2.1 Data

We have constructed a unique dataset from several sources which contains information on every wind farm developed in Denmark and Germany from 1977 to 2005. The data include the location of each project, the number of turbines, the total megawatt capacity, the date of grid-connection, manufacturer identity, and other turbine characteristics, such as rotor diameter and tower heights. We match the project data with the location of each manufacturer's primary production facility, which enables the calculation of road-distances from each manufacturer to each project. This provides us with a spatial source of variation in manufacturer costs which aids in identifying the sources of market segmentation. A key missing variable in our data set is transaction price, which necessitates the use of our model to derive price predictions from first order conditions on profit maximization.[8] Rather than infer border costs through price differences, we use differences in the level of trade; the dependent variable for our analysis is the identity of the manufacturer chosen to supply each project. Appendix A provides a detailed description of the data.

[8]As in most business-to-business industries, transaction prices are confidential. Some firms do publish list prices, which we have collected from industry publications. These prices, however, do not correspond to relevant final prices due to site-specific delivery and installation costs.

In this paper, we concentrate on the period from 1995 to 1996.[9] This has several advantages. First, the set of firms was stable during this time period. There are several medium-to-large firms competing in the market. In 1997, a merger and acquisition wave began, which lasted until 2005. The merger wave, including cross-border mergers, would complicate our analysis of the border effect. Second, site owners in this period were typically independent producers. This contrasts with later periods when utility companies became significant purchasers of wind turbines, leading to more concerns about repeated interaction between purchasers and manufacturers. Third, this period contains several well-established firms and the national price subsidies for wind electricity generation had been in place for several years. Prior to the mid-1990s, the market could be considered an "infant industry" with substantial uncertainty about the viability of firms and downstream subsidies. Fourth, the Danish onshore market saturates after the late 1990s, leaving us with too little variation at that side of the border.[10]

In focusing on a two-year period, we abstract away from some dynamic considerations. Although this greatly simplifies the analysis, it comes with some drawbacks. Most important is that one cannot distinguish sunk costs from fixed costs of entering the foreign export market (Roberts and Tybout, 1997; Das, Roberts, and Tybout, 2007). Because of the small number of firms and the lack of substaintial entry and exit, it would not be possible to reliably estimate sunk costs and fixed costs separately in any case. Instead, we model the decision to enter a foreign market as a one-shot game. This decision does not affect the consistency of our variable cost estimates, whereas our counterfactuals removing fixed costs should be interpreted as removing both sunk and fixed costs. We also abstract away from dynamic effects on production technologies, such as learning-by-doing (see Benkard, 2004). Learning-by-doing would provide firms with an incentive to lower prices below a static profit maximizing level in return for anticipated dynamic gains.[11] Learning-by-doing is less of a concern for the mid-1990s than for earlier years. By 1995, the industry has matured to the extent that it is reasonable to assume that firms were setting prices to maximize expected profits from the sale.

Table 1 displays the market shares or the largest five Danish and German firms in both countries. We take these firms to be the set of manufacturers in our study. All other firms had domestic market shares below 2 percent, no long-term presence in their respective markets, and did not export. In our model, we treat these small turbine producers as a competitive fringe. The German and Danish wind turbine markets were relatively independent from the rest of the world. There was only one firm exporting from outside Germany and Denmark: A Dutch firm, Lagerwey, which sold to 21 projects in Germany (2.26 percent market share) and had a short presence in the German market. We include Lagerwey as part of the competitive fringe.

[9] Appendix A.4 shows that the evidence on market shares and the border effect is stable in subsequent time periods.

[10] Moreover, after the 1990s a substantial fraction of wind turbine installations are offshore, so road-distance to the turbine location is less useful as a source of variation in production costs.

[11] In some cases, this could even lead firms to sell products below cost. See Besanko, Doraszelski, Kryukov, and Satterthwaite (2010) for a fully dynamic computational model of price-setting under learning-by-doing.

Table 1: Major Danish and German Manufacturers

Manufacturer	Nationality	% Market share in Denmark	% Market share in Germany
Vestas	(DK)	45.45	12.04
Micon	(DK)	19.19	8.17
Bonus	(DK)	12.12	5.05
Nordtank	(DK)	11.45	4.73
WindWorld	(DK)	4.38	2.73
Total		92.59	32.72
Enercon	(DE)		32.58
Tacke	(DE)		14.95
Nordex	(DE)	1.68	7.53
Suedwind	(DE)		2.37
Fuhrlaender	(DE)		2.15
Total		94.27	92.3

Notes: Market shares in terms of number of projects installed in 1995-1996. Shares are very similar when projects are weighted by megawatt size.

2.2 Preliminary Analysis of the Border Effect

Table 1 and Figure 2 clearly suggest some degree of market segmentation between Germany and Denmark. Four out of five large German firms—including the German market leader, Enercon—do not have any foreign presence. That all Danish firms enter Germany whereas only one German firm competes in Denmark is consistent with the existence of large fixed costs for exporting. Because the German market is much larger than the Danish market (930 projects were installed in Germany in this period, versus 296 in Denmark—see the map of projects in Figure 2), these fixed costs can be amortized over a larger number of projects in Germany.

For those firms that do export, the decline in market share by moving from foreign to domestic markets may have many different causes. First, market structure changes as the set of firms competing in Denmark is smaller than that in Germany. Second, due to transportation costs, foreign firms will have higher costs than domestic ones simply because projects are likely to be nearer to domestic manufacturing plants. Finally, there may be some variable border costs, which must be paid for each foreign project produced.

We start by exploring the effect of distance as a potential source of market segmentation. The impact of distance on firm costs is illustrated by Figure 4. This figure documents Vestas's declining market share as the distance from its main manufacturing location increases. Whereas Figure 4 suggests that costs increase with distance from the manufacturing base, it cannot easily be used to estimate distance costs. The impact of the border—roughly 160 kilometers from Vestas's manufacturing plant—confounds the relationship. Moreover, in an oligopolistic industry, Vestas's share is a function of not only its own costs but

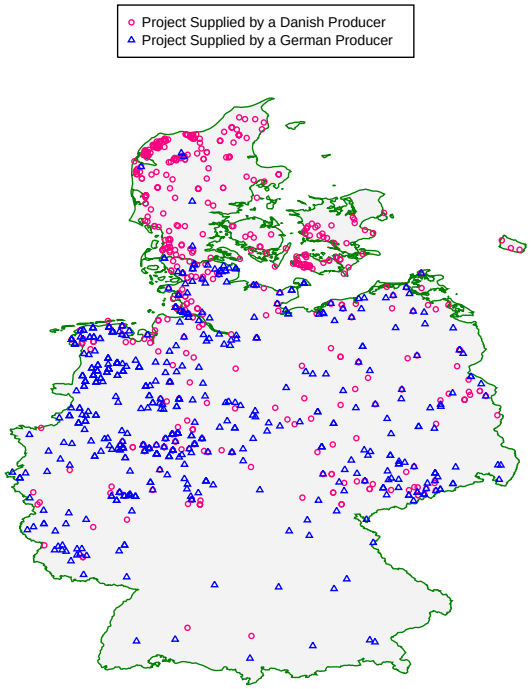


Figure 2: Project Locations

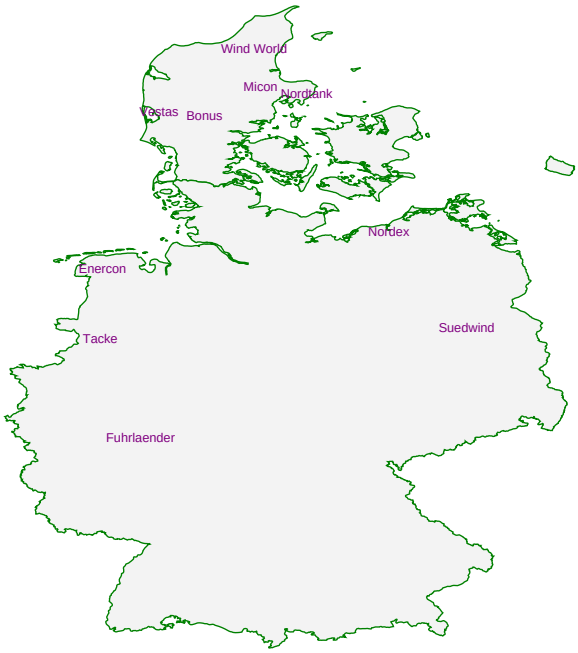


Figure 3: Producer Locations

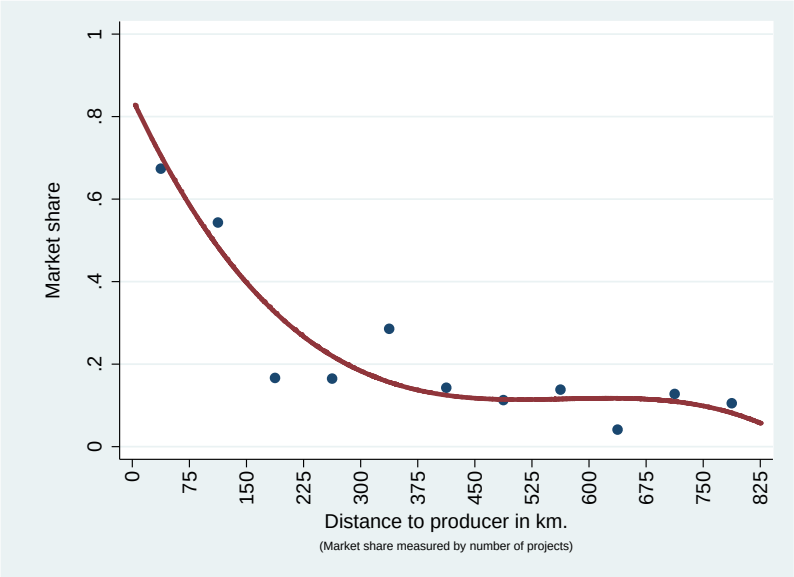
also those of competitor firms. Our model will jointly solve for the probability that each competing firm wins a project based on the project's location in relation to all firms. We are thus able to use the rich variation in projects across space to estimate the impact of distance on firm costs.

We next employ a regression discontinuity design (RDD) to quantify the effect of the border on large Danish firms' market share. Given that wind and demand conditions do not change abruptly, the RDD uncovers the impact of the border. To implement this, we regress a project-level binary variable that takes the value one if it is supplied by a large Danish firms and zero otherwise, to a cubic polynomial of distance from the project to the border, a Germany dummy (to capture the border effect), and interaction terms (see Appendix A.4 for details). Figure 5 plots the fit of this regression. The border dummy is a statistically significant -0.295, which is reflected in the sharp drop in the market share of the largest five Danish firms from around 90 to 60 percent at the border.[12]

These results give us reason to believe that the border matters in the wind turbine industry. Nevertheless, the discontinuity at the border does not separately identify the effect of changes in market structure between Germany and Denmark from the impact of variable border costs. Because variable border costs are incurred precisely at the point where market structure changes, we are unable to use the RDD approach to

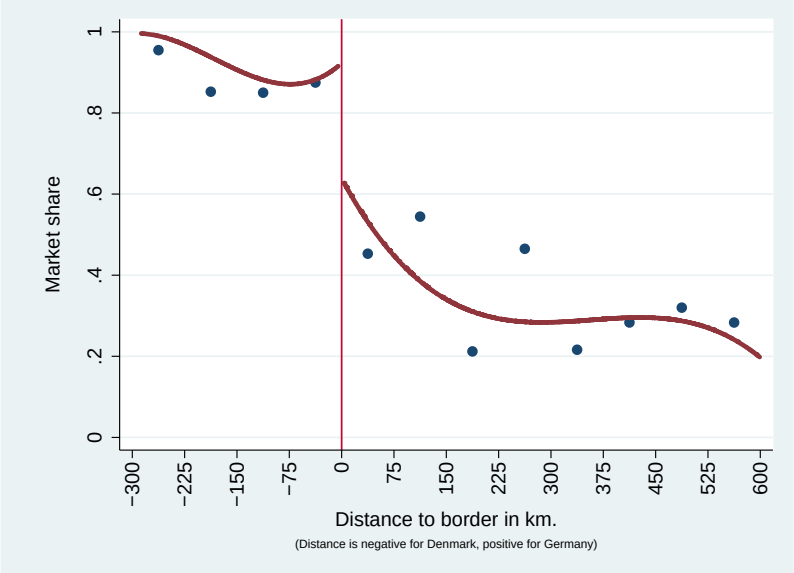
[12]These results are robust to considering projects within various bandwidths of the border, as is standard within the RDD framework. For expository purposes, Figure 5 includes projects in the [-300,600] band.

Figure 4: Market Share of Vestas by Proximity to Primary Production Facility



Notes: Proportion of projects won by Vestas projected on a cubic polynomial of distance to Vestas's production facility. Dots are aggregated market shares in bins of 75 km width.

Figure 5: Market Share of Danish Firms by Proximity to the Border



Notes: Regression discontinuity fit of projects supplied by five large Danish firms using a cubic polynomial of distance to border, a Germany dummy and interaction terms. Regression details are in Appendix A.4. Dots are aggregated market shares in bins of 75 km width.

separate the two effects. This motivates our use of a structural model. The following section proposes and estimates a model to account for the changes market structure at the border by treating the competition for projects as a Bertrand-Nash game.

3 Model

We begin by describing the environment. Denmark and Germany are indexed by $\ell \in \{D, G\}$. Each country has a discrete finite set of large domestic firms denoted by $\mathcal{M}_\ell$ and a local fringe. The number of large domestic firms is equal to the cardinality of this set, $|\mathcal{M}_\ell|$. Large firms are heterogeneous in their location and productivity. There is a fixed number of N_ℓ projects in each country, and they are characterized by their location and size (total megawatt generation capacity). Cross-border competition takes place in two stages: In the first stage, large firms decide whether or not to pay a fixed cost and enter the foreign market. In the second stage, firms bid for all projects in the markets they compete in (they do so in their domestic market by default). Project owners independently choose a turbine supplier among competing firms. We now present the two stages following backward induction, starting with the bidding game.

3.1 Project Bidding Game

In this stage, active firms offer a separate price to each project manager, and project managers choose the offer that maximizes their valuation. The set of active firms is taken as given by all players, as it was realized in the entry stage. For ease of notation, we drop the country index for the moment and describe the project bidding game in one country. The set of active, large firms (denoted by $\mathcal{J}$) and the competitive fringe compete over N projects. $\mathcal{J}$ contains all domestic and foreign firms—if there are any—that entered the market in the first stage, so $\mathcal{M} \subseteq \mathcal{J}$.

The per-megawatt payoff function of a project owner i for choosing firm j is

$$V_{ij} = d_j - p_{ij} + \epsilon_{ij}.$$

The return to the project owner depends on the quality of the wind turbine, d_j, the per-megawatt price p_{ij} charged by manufacturer j, and an idiosyncratic choice-specific shock ϵ_{ij}.[13] It is well known that discrete choice models only identify relative differences in valuations. We thus model a non-strategic fringe as an

[13] We assume away project-level economies of scale by making price bids per-megawatt. Our data does not enable us to identify project-level economies of scale. We check whether foreign turbine manufacturers tend to specialize on larger projects abroad. We find that the average project size abroad is very similar to the average project size at home for each exporting firm.

outside option. We denote it as firm 0 and normalize the return as

$$V_{i0} = \epsilon_{i0}.$$

We assume ϵ_{ij} is distributed i.i.d. across projects and firms according to the Type-I extreme value distribution.[14] The $\boldsymbol{\epsilon}_i$ vector is private information to owners who collect project-specific price bids from producers. The assumption that $\boldsymbol{\epsilon}_i$ is i.i.d and private knowledge abstracts away from the presence of unobservables, which are known to the firms at the time they choose prices but are unknown to the econometrician.[15] After receiving all price bids, denoted by the vector $\mathbf{p}_i$, owners choose the firm that delivers them the highest payoff. Using the familiar logit formula, the probability that owner i chooses firm j is given by

$$Pr[i \text{ chooses } j] \equiv \rho_{ij}(\mathbf{p}_i) = \frac{\exp(d_j - p_{ij})}{1 + \sum_{k=1}^{|\mathcal{J}|} \exp(d_k - p_{ik})} \qquad \text{for } j \in \mathcal{J}. \tag{1}$$

The probability of choosing the fringe is

$$Pr[i \text{ chooses the fringe}] \equiv \rho_{i0}(\mathbf{p}_i) = 1 - \sum_{j=1}^{|\mathcal{J}|} \rho_{ij}(\mathbf{p}_i).$$

Now we turn to the problem of the firm. The cost for firm j to supply project i is a function of its heterogeneous production cost ϕ_j, its distance to the project, and whether or not it is a foreign producer:

$$c_{ij} = \phi_j + \beta_d \cdot \text{distance}_{ij} + \beta_b \cdot \text{border}_{ij}, \tag{2}$$

where

$$\text{border}_{ij} = \begin{cases} 0 & \text{if both } i \text{ and } j \text{ are located in the same country,} \\ 1 & \text{otherwise.} \end{cases}$$

In other words, all firms pay the distance related cost ($\beta_d \cdot \text{distance}_{ij}$), but only foreign firms pay the variable border cost ($\beta_b \cdot \text{border}_{ij}$). The distance cost captures not only the cost of transportation but also serves as a proxy for the cost of post-sale services (such as maintenance), installing remote controllers to monitor wind farm operations, gathering information about sites further away from the manufacturer's location, and maintaining relationships with local contractors who construct turbine towers. The border component captures additional variable costs faced by foreign manufacturers. This may include the cost of dealing with

[14] Project owners do not have any home bias in the sense that ϵ_{ij}'s are drawn from the same distribution for all producers in both countries.

[15] For example, if local politics or geography favors one firm over another in a particular region, firms would account for this in their pricing strategies, but we are unable to account for this since this effect is unobserved to us. In Appendix B, we address the robustness of our estimate to local unobservables of this type.

project approval procedures in the foreign market and coordinating transportation of bulky components with various national and local agencies.

Firms engage in Bertrand competition by submitting price bids for projects in the markets in which they are active.[16] They observe the identities and all characteristics of their competitors (i.e., their quality and marginal cost for each project) except the valuation vector $\boldsymbol{\epsilon}_i$. The second stage is thus a static game with imperfect, but symmetric, information. In a pure-strategy Bayesian-Nash equilibrium, each firm chooses its price to maximize expected profits given the prices of other firms:[17]

$$E[\pi_{ij}] = \max_{p_{ij}} \ \rho_{ij}(p_{ij}, \mathbf{p}_{i,-j}) \cdot (p_{ij} - c_{ij}) \cdot S_i,$$

where S_i is the size of the project in KW. The first order condition reads as follows:

$$\begin{aligned} 0 &= \frac{\partial \rho_{ij}(p_{ij}, \mathbf{p}_{i,-j})}{\partial p_{ij}}(p_{ij} - c_{ij}) + \rho_{ij}(p_{ij}, \mathbf{p}_{i,-j}), \\ p_{ij} &= c_{ij} - \frac{\rho_{ij}(p_{ij}, \mathbf{p}_{i,-j})}{\partial \rho_{ij}(p_{ij}, \mathbf{p}_{i,-j})/\partial p_{ij}}. \end{aligned}$$

Exploiting the properties of the logit form, this expression simplifies to an optimal mark-up pricing condition:

$$p_{ij} = c_{ij} + \frac{1}{1 - \rho_{ij}(p_{ij}, \mathbf{p}_{i,-j})}. \tag{3}$$

The mark-up is increasing in the (endogenous) probability of winning the project and is thus a function of the set of the firms active in the market and their characteristics. Substituting (3) into (1), we get a fixed-point problem with $|\mathcal{J}|$ unknowns and $|\mathcal{J}|$ equations for each project i:

$$\rho_{ij} = \frac{\exp\left(d_j - c_{ij} - \frac{1}{1-\rho_{ij}}\right)}{1 + \sum_{k=1}^{|\mathcal{J}|} \exp\left(d_k - c_{ik} - \frac{1}{1-\rho_{ik}}\right)} \qquad \text{for } j \in \mathcal{J}. \tag{4}$$

Our framework fits into the class of games for which Caplin and Nalebuff (1991) show the existence of a unique pure-strategy equilibrium. Using the optimal mark-up pricing condition, the expected profits of manufacturer j for project i can be calculated as,

$$E[\pi_{ij}] = \frac{\rho_{ij}}{1 - \rho_{ij}} S_i.$$

[16] Industry experts we interviewed indicated that there was an excess supply of production capacity in the market during these years. One indication of this is that many firms suffered from low profitability, sparking a merger wave. Therefore, it is not likely that capacity constraints were binding in this period.

[17] We assume that firms are maximizing expected profits on a project-by-project level. These abstracts away from economics of density in project locations–i.e., the possibility that by having several projects close together they could be produced and maintained at a lower cost. We are address the robustness of our model to the presence of economies of density in Appendix B.

Potential exporters take expected profits into account in their entry decisions. We turn to the entry game in the next section.

3.2 Entry Game

Before bidding on projects, an entry stage is played in which all large firms simultaneously decide whether or not to be active in the foreign market by incurring a firm-specific fixed cost f_j. This captures expenses that can be amortized across all foreign projects, such as establishing a foreign sales office, gaining regulatory approvals, or developing the operating software satisfying the requirements set by national grids.

Let $\Pi_j(\mathcal{J}_{-j} \cup j)$ be the expected profit of manufacturer j in the foreign market where $\mathcal{J}_{-j}$ is the set of active bidders other than j. This is simply the sum of the expected profit of bidding for all foreign projects:

$$\Pi_j(\mathcal{J}_{-j} \cup j) = \sum_{i=1}^{N} E[\pi_{ij}(\mathcal{J}_{-j} \cup j)]. \tag{5}$$

Manufacturer j enters the foreign market if its expected return is higher than its fixed cost:

$$\Pi_j(\mathcal{J}_{-j} \cup j) \geq f_j. \tag{6}$$

Note that this entry game may have multiple equilibria. Following the literature initiated by Bresnahan and Reiss (1991), we assume that the observed decisions of firms are the outcome of a pure-strategy equilibrium; therefore, if a firm in our data is active in the foreign market, (6) must hold for that firm. On the other hand, if firm j is not observed in the foreign market, one we can infer the following lower bound on fixed export cost:

$$\Pi_j(\mathcal{J}_{-j} \cup j) \leq f_j. \tag{7}$$

We use these two necessary conditions to construct inequalities that bound f_j from above or from below by using the estimates from the bidding game to impute the expected payoff estimates of every firm for any set of active participants in the foreign market.[18] We now turn to the estimation of the model.

4 Estimation

Estimation proceeds in two steps: In the first step, we estimate the structural parameters of the project-bidding game. In the second step, we use these estimates to solve for equilibria in the project-bidding game

[18] Several papers (e.g., Pakes, Porter, Ho, and Ishii, 2006; Ciliberto and Tamer, 2009) proposed using bounds to construct moment inequalities for use in estimating structural parameters. Holmes (2011) and Morales, Sheu, and Zahler (2011) applied this methodology to the context of spatial entry and trade. Because we observe only a single observation of each firm's entry decision, a moment inequality approach is not applicable in our setting.

with counterfactual sets of active firms to construct the fixed costs bounds. Before proceeding with the estimation, we must define the set of active firms in every country. Under our model, the set of firms that have positive sales in a country is a consistent estimate of the active set of firms; therefore, we define a firm as active in the foreign market if it has any positive sales there.[19]

We now reintroduce the country index: ρ_{ij}^{ℓ} is firm j's probability of winning project i in country ℓ. The number of active firms in market ℓ is $|\mathcal{J}_\ell|$, and $\text{border}_{ij}^{\ell}$ equals zero if project i and firm j are both located in country ℓ and one otherwise. Substituting the cost function (2) into the winning probability (4), we get

$$\rho_{ij}^{\ell} = \frac{\exp\left(d_j - \phi_j - \beta_d \cdot \text{distance}_{ij} - \beta_b \cdot \text{border}_{ij}^{\ell} - \frac{1}{1-\rho_{ij}^{\ell}}\right)}{1 + \sum_{k=1}^{|\mathcal{J}_\ell|} \exp\left(d_k - \phi_k - \beta_d \cdot \text{distance}_{ik} - \ \beta_b \cdot \text{border}_{ik}^{\ell} - \frac{1}{1-\rho_{ik}^{\ell}}\right)}. \tag{8}$$

From this equation, one can see that firms' production costs ϕ_j and quality level d_j are not separately identified given our data.[20] We thus jointly capture these two effects by firm fixed-effects $\xi_j = d_j - \phi_j$.

We collect the parameters to estimate into the vector $\theta = (\beta_b, \beta_d, \xi_1, \ldots, \xi_{|\mathcal{M}_D|+|\mathcal{M}_G|})$. We estimate the model via constrained maximum likelihood, where the likelihood of the data is maximized subject to our equilibrium constraints. The likelihood function of the project data has the following form:

$$L(\rho) = \prod_{\ell \in \{D,G\}} \prod_{i=1}^{N_\ell} \prod_{j=1}^{|\mathcal{J}_\ell|} \left(\rho_{ij}^{\ell}\right)^{y_{ij}^{\ell}}, \tag{9}$$

where $y_{ij}^{\ell} = 1$ if manufacturer j is chosen to supply project i in country ℓ and 0 otherwise. $\hat{\theta}$, together with the vector of expected project win probabilities $\hat{\rho}$, solves the following problem:

$$\begin{aligned} \max_{\theta,\, \rho} \quad & L(\rho) \\ \text{subject to:} \quad & \rho_{ij}^{\ell} = \frac{\exp\left(\xi_j - \beta_d \cdot \text{distance}_{ij} - \beta_b \cdot \text{border}_{ij}^{\ell} - \frac{1}{1-\rho_{ij}^{\ell}}\right)}{1 + \sum_{k=1}^{|\mathcal{J}_\ell|} \exp\left(\xi_k - \beta_d \cdot \text{distance}_{ik} - \ \beta_b \cdot \text{border}_{ik}^{\ell} - \frac{1}{1-\rho_{ik}^{\ell}}\right)} \\ & \text{for } \ell \in \{D, G\},\ i \in \{1, ..., N_\ell\},\ j \in \mathcal{J}. \end{aligned} \tag{10}$$

Our estimation is an implementation of the Mathematical Programming with Equilibrium Constraints (MPEC) procedure proposed by Judd and Su (2011). They show that the estimator is equivalent

[19] Note that every active firm has a positive probability of winning every project. As the number of projects goes to infinity, every active firm wins at least one project. We thus consider firms with zero sales in a market as not having entered in the first stage and exclude them from the set of active firms there.

[20] The difference between productivity and quality would be identified if we had data on transaction prices. Intuitively, for two manufacturers with similar market shares, high prices would be indicative of higher quality products while low prices would be indicative of lower costs.

to a nested fixed-point estimator in which the inner loop solves for the equilibrium of all markets, and the outer loop searches over parameters to maximize the likelihood. The estimator therefore inherits all the statistical properties of a fixed-point estimator. It is consistent and asymptotically normal as the number of projects tends to infinity. For the empirical implementation, we reformulate the system of constraints in (10) in order to simplify its Jacobian. In our baseline specification, this is a problem with 12,314 variables (12 structural parameters and 12,302 equilibrium win probabilities for all firms competing for each project) and 12,302 equality constraints. We describe the details of the computational procedure in Appendix C.

As a robustness check on our baseline specification, we also try an alternative cost specification in which distance related costs are firm-specific:

$$c_{ij}^{\ell} = \phi_j + \beta_{dj} \cdot \text{distance}_{ij} + \beta_b \cdot \text{border}_{ij}^{\ell}.$$

Note that the difference between this and the baseline specification (2) is that distance cost coefficients are heterogeneous (β_{dj} vs. β_d). This cost function is consistent with Holmes and Stevens (2012), who document that in U.S. data large firms tend to ship further away, even when done domestically.[21] If heterogeneous shipping costs were present in the wind turbine industry, they might bias our baseline estimate of the border effect upward through a misspecification of distance costs, since smaller firms would not export due to higher transport costs instead of the border effect. In the following section, we present results for both specifications.

Once the structural parameters are recovered, one can calculate bounds on the fixed costs of entry for each firm, f_j, using the equations (6) and (7). This involves resolving the model with the appropriate set of firms while holding the structural parameters fixed at their estimated values. Finally, a parametric bootstrap procedure helps to calculate the standard errors for these bounds.[22]

4.1 Parameter Estimates

Estimation results are presented in Table 2, with the baseline specification reported in the first column. Both variable costs are economically and statistically significant. Based on our estimate, the cost of supplying a foreign project is equivalent to an additional 432 kilometers of distance between the manufacturing location and the project site ($\beta_b/\beta_d = 0.432$). The mean distance from Danish firms to German projects in our data is 623 kilometers; the distance from German firms to Danish projects is 602 kilometers. As a consequence, border frictions represent roughly 40 percent of exporters' total delivery costs.

[21] They rationalize this observation in a model where heterogeneous firms invest in their distribution networks. Productive firms endogenously face a lower "iceberg transportation cost."

[22] To be specific, we repeatedly draw θ_b from the asymptotic distribution of $\hat{\theta}$ and recalculate the bound each time. Under the assumptions of the model, the distribution of bound statistic generated by this procedure is a consistent estimate of the true distribution.

To get a sense of the importance of distance-related costs on market outcomes, we calculate the distance elasticity of the equilibrium probability of winning a project for every project-firm combination.[23] For exporters, the median distance elasticity ranges from .95 to 1.40. That is, the median effect of a one percent increase in the distance from an exporting firm to a project abroad (holding all other firms' distances constant) is a decline of .95 to 1.40 percent in the probability of winning the project. For domestic firms, the median distance elasticities are lower, ranging from .17 to .83. The difference is due to both the smaller distances firms must typically travel to reach domestic projects and the impact of the border on equilibrium outcomes. It appears that distance costs have a significant impact on firm costs and market shares for both foreign and domestic firms.

As discussed above, the firm fixed effects reflect the combination of differences in quality and productivity across firms. We find significant differences between them. It is not surprising that the largest firms, Vestas and Enercon, have the highest fixed effects. Although there is significant within-country dispersion, Danish firms generally appear to be stronger than German ones. The results suggest that controlling for firm heterogeneity is important for correctly estimating border and distance costs.

Since our model delivers expected purchase probabilities for each firm at each project site, we can use the regression discontinuity approach to visualize how well our model fits the observed data. Figure 6 presents this comparison. The horizontal axis is the distance to the Danish-German border, where negative distance is inside Denmark. The red (solid) is the raw data fit. This is the same curve as that presented in Figure 5, relating distance-to border and a border dummy to the probability of a Danish firm winning a project. In particular, this regression does not control for project-to-firm distances. The blue (dotted) curve is fitted using the expected win probabilities calculated from the structural model. These probabilities depend explicitly on our estimates of both firm heterogeneity and project-to-manufacturer distances but do not explicitly depend on distance to the border (as this indirectly affects costs for firms in our model). Note that predicted win probabilities are nonlinear despite the linearity of costs. This is due to the nonlinear nature of the model as well as the rich spatial variation of mark-ups predicted by the model. The size of the discontinuity is somewhat larger using the structural model, although the qualitative result that the border effect is large is apparent using both methods. Overall, the model fits the data well.

To address our concern that differences in distance costs across firms may affect our estimation of the border effect, we allow for heterogeneity in distance costs in the second column of Table 2. The border variable cost coefficient is practically unchanged and remains strongly significant, indicating that our

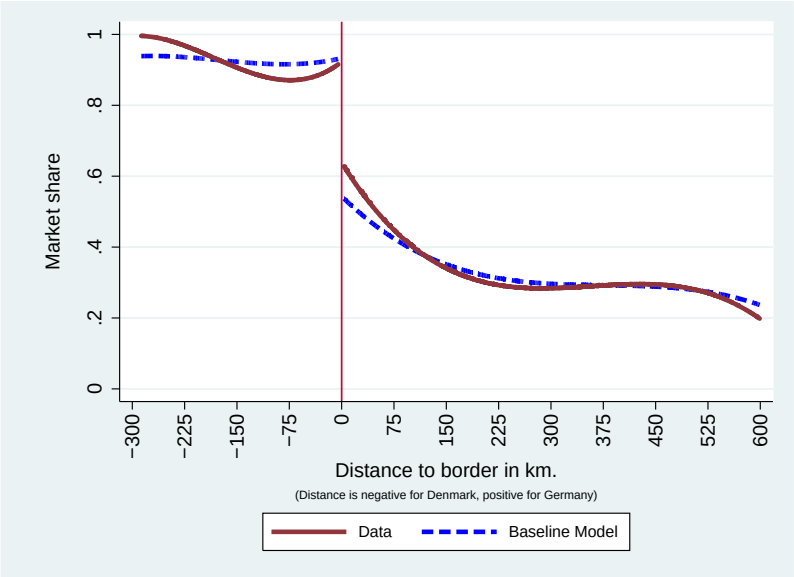
[23]The distance elasticities we report are a function of the characteristics of all firms at a particular project site in a very specific industry. It is difficult to directly compare these distance elasticities with distance elasticities of aggregated trade volumes frequently reported in the trade literature that rely on national or regional capital distance proxies (e.g., McCallum (1995); Eaton and Kortum (2002); Anderson and van Wincoop (2003))

Table 2: Maximum Likelihood Estimates

	Baseline	Heterogeneous Distance Costs
Border Variable Cost, β_b	0.869	0.867
	(0.219)	(0.239)
Distance Cost (100km), β_d	0.201	
	(0.032)	
Bonus (DK)		0.169
		(0.066)
Nordtank (DK)		0.277
		(0.073)
Micon (DK)		0.134
		(0.051)
Vestas (DK)		0.287
		(0.049)
WindWorld (DK)		0.016
		(0.068)
Enercon (DE)		0.296
		(0.063)
Fuhrlaender (DE)		1.794
		(0.236)
Nordex (DE)		-0.071
		(0.087)
Suedwind (DE)		-0.231
		(0.104)
Tacke (DE)		0.103
		(0.071)
Firm Fixed Effects, ξ_j		
Bonus (DK)	2.473	2.332
	(0.223)	(0.297)
Nordtank (DK)	2.526	2.811
	(0.229)	(0.326)
Micon (DK)	3.097	2.786
	(0.221)	(0.268)
Vestas (DK)	3.805	4.180
	(0.215)	(0.265)
WindWorld (DK)	1.735	0.818
	(0.273)	(0.418)
Enercon (DE)	3.533	3.859
	(0.175)	(0.270)
Fuhrlaender (DE)	0.330	3.305
	(0.263)	(0.506)
Nordex (DE)	1.782	0.683
	(0.203)	(0.400)
Suedwind (DE)	0.537	-1.188
	(0.270)	(0.510)
Tacke (DE)	2.389	2.104
	(0.177)	(0.263)
Log-Likelihood	-2363.00	-2315.82
N	1226	1226

Notes: Standard errors in parentheses.

Figure 6: Model Fit: Expected Danish Market Share by Distance to the Border



Notes: Data line is the same as in Figure 5. The model line is the regression discontinuity fit of probability of winning each project by Danish firms on a cubic polynomial of distance to border, Germany dummy, and interaction terms.

border effect estimate is not driven by distance cost heterogeneity. Turning to the distance costs themselves, most firms, particularly the larger ones, have distance costs that are close to our homogenous distance cost estimate. It does not appear that small firms have systematically higher distance costs. The smallest firm in our data, Suedwind, is estimated to be distance loving; this firm is based in Berlin, but has built several turbines in the west of Germany.[24] While a formal likelihood ratio test rejects the null hypothesis of homogeneous distance costs, the estimation results indicate that heterogeneous distance costs are not driving cross-border differences in this industry. Therefore, we use our baseline specification for the counterfactual analysis below.

4.2 Fixed Cost Bounds

Not all firms enter the foreign market; rather, firms optimally choose whether or not to export by weighing their fixed costs of entry against the expected profits from exporting. Hence, firm-level heterogeneity in operating profits, fixed costs, or both is necessary to rationalize the fact that different firms make different exporting decisions.[25] Since our model naturally allows for heterogeneity in firm operating profits, this section considers whether heterogeneity in firms' fixed costs of exporting are also needed to rationalize observed entry decisions.

[24] Nordex, who is also located in the east of Germany, also has a negative coefficient, but it is statistically insignificant.

[25] The canonical Melitz (2003) model assumes homogenous fixed costs and heterogeneity in operating profits. Eaton, Kortum, and Kramarz (2011) show that heterogeneity in fixed costs is also necessary to fit the export patterns in French firm-level data.

Table 3: Export Fixed Cost Bounds (f_j)

	Lower	Upper
Bonus (DK)		47.55
		(19.52)
Nordtank (DK)		43.29
		(8.91)
Micon (DK)		80.13
		(13.62)
Vestas (DK)		164.32
		(23.60)
WindWorld (DK)		17.35
		(3.93)
Enercon (DE)	22.32	
	(4.87)	
Fuhrlaender (DE)	0.66	
	(0.32)	
Nordex (DE)		6.33
		(1.82)
Suedwind (DE)	1.26	
	(0.45)	
Tacke (DE)	7.24	
	(1.71)	

Notes: Scale is normalized by the variance of ϵ. Standard errors in parentheses.

Since we only observe a single export decision for each firm, fixed costs are not point identified. Nevertheless, the model helps to place a bound on them. Firms optimally make their export decision based on the level of fixed costs of foreign entry and on the operating profits they expect in the export market as described in Section 3.2. Based on the parameter estimates in Table 2, we can derive counterfactual estimates of expected operating profits for any set of active firms in the Danish and German markets. Therefore, we can construct an upper bound on fixed costs for firms entering the foreign market using (6), and a lower bound on fixed costs for firms that stay out of the foreign market using (7). While the scale of these bounds is normalized by the variance of the extreme-value error term, comparing them across firms gives us some idea of the degree of heterogeneity in fixed costs.

Table 3 presents the estimates of fixed cost bounds for each firm. The intersection of the bounds across all firms is empty. For example, there is no single level of fixed costs that would simultaneously justify WindWorld entering Germany and Enercon not entering Denmark; hence, some heterogeneity in fixed costs is necessary to explain firm entry decisions.

One possibility is that fixed cost for entering Germany differ from those for entering Denmark. Since all Danish firms enter the Danish market, any fixed cost below 17.35 (the expected profits of WindWorld for

entering Germany) would rationalize the observed entry pattern. In Germany however, the lower and upper bound of Enercon and Nordex have no intersection. Some background information about Nordex supports the implication of the model that Nordex may be subject to much lower costs than Enercon to enter into the Danish market. Nordex was launched as a Danish company in 1985 but shifted its center of business and production activity to Germany in the early 1990s. As a consequence, Nordex could keep a foothold in the Danish market at a lower cost than could the other German firms, which would need to form contacts with Danish customers from scratch.[26]

Of course, the Nordex anecdote also highlights some important caveats with regard to our bounds. By assuming a one-shot entry game, we are abstracting away from entry dynamics. If exporting is less costly to continue than to initiate, then the bounds we calculate—which consider only profits from operating in 1995 and 1996—will be biased downward. Data limitations, particularly the small number of firms, prevent us from extending the model to account for dynamic exporting decisions along the lines of Das, Roberts, and Tybout (2007). Nevertheless, our results illustrate the degree of heterogeneity in fixed costs that is necessary to explain entry patterns.[27]

5 Border Frictions, Market Segmentation, and Welfare

We now use the model to study the impact of border frictions on national market shares, firm profits,and consumer welfare. We perform a two-step counterfactual analysis. The first step eliminates fixed costs of exporting, keeping in place variable costs incurred at the border.[28] This counterfactual allows us to examine the importance of the change in the competitive environment at the border. The second step further removes the variable cost of the border by setting β_b equal to zero. This eliminates all border frictions such that the only sources of differing market shares across national boundaries are plant-to-project distances and firm heterogeneity.[29] While the results of this experiment constitute an estimate of what can be achieved if border frictions could be entirely eliminated, it is important to keep in mind that natural barriers, such as different languages, will be difficult to eliminate in practice.

[26] Because of Nordex's connection to Denmark, we perform a robustness check on our variable border cost estimate by re-estimating the model allowing Nordex to sell in Denmark without having to pay the border variable cost. The border cost estimate increases in this specification, but the difference is not statistically significant. Since Nordex is the only exporting German firm, this robustness check also serves as a check on our specification of symmetric border costs. See Balistreri and Hillberry (2007) for a discussion of asymmetric border frictions.

[27] It is important to note that the variable cost estimates presented in Table 2, as well as the counterfactual results below, are robust to dynamic entry as long as firm pricing decisions have no impact on future entry decisions. This assumption is quite common in the literature on structural oligopoly models, e.g., Ericson and Pakes (1995).

[28] We implicitly assume that the change in market structure does not induce domestic firms to exit the industry, or new firms to be created.

[29] We eliminate first fixed border costs and then variable costs because changes in variable border costs when fixed costs are still positive could induce changes in the set of firms that enter foreign markets. Because they are not point identified, we are unable to estimate fixed border costs. Even with reliable estimates, the entry stage with positive fixed costs is likely to result in multiple equilibria.

Table 4: Counterfactual Market Shares of Large Firms (%)

		Data	Baseline Estimates	No Fixed Costs	No Border
Denmark	Danish Firms	92.57	92.65	83.95	74.26
			(1.52)	(2.26)	(3.64)
	German Firms	1.69	2.18	11.56	21.94
			(0.60)	(2.05)	(3.88)
Germany	Danish Firms	32.37	32.42	32.42	49.32
			(5.42)	(5.42)	(7.55)
	German Firms	59.57	59.24	59.24	44.90
			(3.93)	(3.93)	(5.80)

Notes: Market share measured in projects won. Standard errors in parentheses.

5.1 Market Shares and Segmentation

We begin our analysis by considering how national market shares in each country react to the elimination of border frictions. Furthermore, because market shares are directly observed in the data, the baseline model's market share estimates can also be used to assess the fit of our model to national level aggregates. Table 4 presents the market share of the major firms of Denmark and Germany in each country, with the fringe taking the remainder of the market. Comparing the first two columns, the baseline predictions of the model closely correspond to the observed market shares. All of the market shares are within the 95 percent confidence interval of the baseline predictions, which suggests that the model has a good fit.

In the third column, we re-solve the model by eliminating fixed costs of exporting and keeping the variable border cost in place. In response, the four German firms that previously competed only domestically start exporting to Denmark. As a result, the market share of German firms in Denmark rises more than 10 percentage points.[30] Danish firms, however, still maintain a substantial market share advantage in their home market. Since all five large Danish firms already compete in Germany, there is no change in market shares on the German side of the border when fixed costs of exporting are removed.[31] The difference in response to the elimination of fixed costs between the Danish and German markets is obvious, but instructive. The reduction or elimination of border frictions can have very different effects based on market characteristics. In our case, because there are more projects in Germany than in Denmark, the return to entry is much larger in Germany. This may be one reason why we see more Danish firms entering Germany than vice versa.[32] As a result, reducing fixed costs of exporting to Germany has no effect on market outcomes, whereas the

[30] For space and clarity, we do not report standard errors of changes in market shares in Table 4. All of the (non-zero) changes in market shares across counterfactuals are statistically significant.

[31] Because of this duplication, we simply omit the column which removes fixed cost of entry in Germany in the tables below.

[32] This argument assumes fixed costs of exporting are of the same order of magnitude for both countries, which appears to be the case.

impact of eliminating fixed cost of exporting to Denmark is substantial.

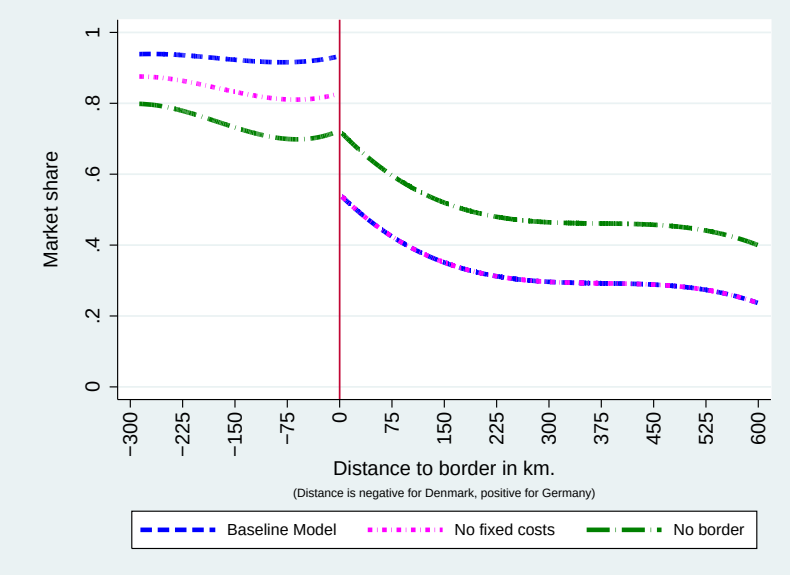
The fourth and final column of Table 4 displays the model prediction of national market shares if the border were entirely eliminated. In addition to setting f_j equal for all firms, we also eliminate variable border costs by setting β_b equal to zero.[33] This results in a large increase in imports on both sides of the border. The domestic market share of Danish firms falls from 92.6 percent to 74.3 percent. The domestic market share of large Danish firms remains high due to firm heterogeneity and the fact that they are closer to Danish projects. In Germany, a slight majority of projects imports Danish turbines once the border is eliminated, which reflects the strength of Danish firms (especially Vestas) in the wind turbine industry. On both sides of the border, we see an approximate 20 percent increase in import share when the national boundary is eliminated.

Overall, our results indicate that border frictions generate significant market segmentation between Denmark and Germany. As a back of the envelope illustration, consider the difference between the market share of Danish firms in the two markets. The gap in the data and baseline model is roughly 60 percentage points. Not all of this gap can be attributed to border frictions since differences in transportation costs due to geography are also responsible for part of the gap. However, when we remove border frictions, our counterfactual analysis indicates that the gap shrinks to 25 percentage points. More than half of the market share gap is thus attributable to border frictions. When considering the sources of border frictions, we find that removing fixed costs of exporting alone accounts for one-third of the market share gap that is attributed to border frictions, while the remaining two-thirds are realized by removing both fixed and variable border frictions. Since fixed and variable costs interact, the overall impact of border frictions cannot be formally decomposed into fixed and variable cost components. We take these results as evidence that both fixed and variable border frictions are substantial sources of market segmentation.

In addition to national market share averages, our model allows us to examine predicted market shares at a particular point in space. Using the RDD approach describe above, Figure 7 visualizes the impact of the counterfactual experiments. The blue (dashed) line represents expected market shares baseline model, and is identical to that presented in Figure 6. The red (dotted) line displays counterfactual expected market shares when fixed border costs are removed. This reduces domestic market share of Danish firms since more German firms enter, but leaves market shares unchanged in Germany since all firms were already competing there. Finally, the green (dashed-dotted) line shows the counterfactual estimates when all border costs are eliminated. The discontinuity at the border is entirely eliminated, and only the impact of firm-to-

[33] Because adjustments to variable costs may result in a change in firms optimal entry decisions, we are unable to perform a counterfactual eliminating variable border costs alone.

Figure 7: Counterfactuals: Expected Danish Market Share by Distance to the Border



Notes: Regression discontinuity fit of projects won by Danish firms on a cubic polynomial of distance to border, Germany dummy, and interaction terms.

manufacturer distances cause differences in market share on either side of the border.[34]

5.2 Firm Profits

We now turn to an analysis of winners and losers from border frictions, starting with individual firms. Table 5 presents the level of operating profits under the baseline and two counterfactual scenarios, calculated according to equation (5). While the scale of these profit figures is arbitrary, they allow for comparison both across firms and across scenarios. The table separates profits accrued in Germany and Denmark for each firm. For example, in the baseline scenario, we see that Bonus made 47.06 in profits in Denmark, and 47.55 in Germany. If the border were removed entirely, Bonus's profits in Denmark would fall to 34.83, while their profits in Germany would rise to 75.46. On overall, Bonus would see its total profits increase as a result of the elimination of border frictions, as gains in Germany would more than offset loses from increased competition in Denmark.

The situation is different for German firms. When fixed costs are eliminated, the large German firms—Enercon and Tacke—take the lion's share of the gains. However, all German firms—even the largest firm, Enercon—loose from the entire elimination of border frictions. Underlying this result is the significant asymmetry in size and productivity between Germany and Denmark. The losses German firms face due to

[34]The kink at the boundary is an artifact of interaction terms in the RDD which implies that we estimate either side of the border as a separate cubic polynomial in distance to the border. The bottom line is that there is no discontinuity at the boundary when all border effects are removed.

Table 5: Baseline and Counterfactual Profit Estimates

	Denmark			Germany	
	Baseline	No Fixed Costs	No Border	Estimates	No Border
Bonus (DK)	47.06	41.02	34.83	47.55	75.46
	(13.00)	(11.98)	(10.71)	(19.52)	(28.88)
Nordtank (DK)	44.70	38.98	33.11	43.29	68.72
	(4.97)	(4.50)	(4.19)	(8.91)	(13.73)
Micon (DK)	82.76	72.63	62.07	80.13	126.74
	(7.36)	(6.80)	(6.75)	(13.62)	(21.14)
Vestas (DK)	156.96	139.46	120.69	164.32	256.08
	(14.60)	(12.46)	(11.83)	(23.60)	(37.23)
WindWorld (DK)	20.73	18.13	15.44	17.35	27.59
	(3.19)	(2.76)	(2.49)	(3.93)	(6.32)
Enercon (DE)		21.46	42.56	428.91	305.06
		(4.54)	(9.37)	(48.68)	(53.60)
Fuhrlaender (DE)		0.57	1.14	17.31	11.98
		(0.26)	(0.56)	(4.20)	(3.28)
Nordex (DE)	6.33	5.43	10.79	75.69	51.24
	(1.82)	(1.48)	(2.45)	(15.15)	(13.20)
Suedwind (DE)		1.09	2.16	21.74	14.85
		(0.37)	(0.78)	(5.23)	(3.90)
Tacke (DE)		6.47	12.93	151.86	104.83
		(1.42)	(3.19)	(16.60)	(17.33)

Notes: Scale is normalized by variance of ϵ. Standard errors in parentheses.

increased competition in the larger German market overwhelm all gains they receive from frictionless access to the Danish market. Our model estimates Danish firms to be highly productive, so eliminating the border is quite costly to German incumbents. In addition, variable border frictions are estimated to be so high that even a small Danish exporter like WindWorld becomes much more competitive in Germany when they are removed. Despite being a relatively small player, WindWorld gains from the elimination of border frictions since increased profits in the larger Germany market outweigh its losses at home. However, WindWorld's gains are insignificant when compared to the gains of the large Danish firms, such as Vestas. Overall, we find that because a German firm's domestic market is considerably larger than its export market, border frictions protect the profit of German firms over those of Danish firms.

5.3 Consumer Surplus and Welfare

We now analyze the overall impact of the border on welfare in the Danish and German wind turbine markets. For each country, Table 6 presents consumer surplus (i.e., surplus accruing to site owners) and firm profits (aggregated by country) under the baseline and our two counterfactual scenarios. The relative changes in

Table 6: Counterfactual Welfare Analysis by Country

		Baseline	No Fixed Costs		No Border	
		(Levels)	(Levels)	(% Chg)	(Levels)	(% Chg)
Denmark	(A) Consumer Surplus	70.15	73.46	4.72	77.42	10.36
		(4.94)	(4.97)	(1.03)	(5.38)	(2.19)
	(B) Danish Firm Profits	29.33	25.83	-11.92	22.16	-24.44
		(0.54)	(0.74)	(2.26)	(1.26)	(4.47)
	(C) German Firm Profits	0.53	2.91	452.99	5.79	999.18
		(0.15)	(0.55)	(122.97)	(1.13)	(297.29)
	Domestic Surplus (A+B)	99.47	99.29	-0.18	99.58	0.10
		(5.17)	(5.11)	(0.07)	(5.09)	(0.25)
	Total Surplus (A+B+C)	100.00	102.21	2.21	105.37	5.37
		(5.09)	(5.07)	(0.51)	(5.39)	(1.28)
Germany	(A) Consumer Surplus	68.99			79.57	15.34
		(6.42)			(8.30)	(1.90)
	(B) Danish Firm Profits	10.43			16.41	57.27
		(1.59)			(2.41)	(4.96)
	(C) German Firm Profits	20.58			14.44	-29.84
		(1.86)			(2.31)	(5.62)
	Domestic Surplus (A+C)	89.57			94.01	4.96
		(5.78)			(6.68)	(1.39)
	Total Surplus (A+B+C)	100.00			110.42	10.42
		(6.72)			(8.59)	(1.77)

Notes: Levels are scaled such that baseline total surplus from projects within a country is 100. "% Chg" is percent change from baseline level. Standard errors in parentheses.

consumer surplus across scenarios are invariant to the scale of ϵ, so we normalize the consumer surplus in the baseline scenario to 100 for expositional ease.[35] We define domestic surplus as the total surplus in the country that accrues to consumers and domestic firms.

The first column reports the breakdown of surplus under the baseline scenario, we see that in both Denmark and Germany, consumers receive roughly 70 percent of the total surplus. In Denmark, the bulk of the remaining 30 percent goes to Danish firms (recall that only one German firm is active in Denmark), while in Germany, approximately 10 percent goes to Danish firms and 20 percent to German firms.

The next two columns present results from the counterfactual where only fixed costs of entry are removed. As discussed above, this counterfactual only affects the Danish market outcomes, since all Danish firms already sell in Germany in the baseline scenario. We report both surplus levels, and the percentage change from the baseline level. Note that, because of the correlation in the level estimates due to the uncertainty in firm fixed effects, the percent change estimates are much more precise than a naïve comparison of the level estimates would suggest. Removing fixed costs of exporting causes four German firms to enter

[35] Because of its larger size, the total surplus in Germany is much larger than in Denmark, cross country comparisons of total surplus are available by request.

the Danish market, which both increases price competition and provides additional variety to Danish site owners. As a result, consumer surplus increases by 4.72 percent. Danish firms, facing harsher domestic competition, see profits decline by 11.92 percent. Since the number of German firms increased from one to five, total German profits skyrocket in percentage terms, however this is due to a very small initial base. Even after removing fixed costs, German firms take less than three percent of the available surplus in Denmark in profits. The gains of Danish consumers from removing fixed export costs are almost perfectly offset by the loses from Danish firms. Domestic surplus actually declines by a statistically significant but economically negligible amount. When we account for the gains by German firms, total surplus increases by the statistically and economically significant 2.21 percent.

The final two columns of Table 6 display the welfare effects of removing border frictions entirely. As we would expect, site owners see significant benefits, and consumer surplus rises by 10.36 percent in Denmark and 15.34 percent in Germany. The increase in Denmark is more than twice as high as the increase realized from only removing fixed border costs. These increases come at the cost of domestic producers, who see home profits decline by 24.44 percent in Denmark and 29.84 percent in Germany.[36] In Denmark, the removal of border frictions results in a transfer of surplus from domestic firms to consumers, netting to essentially no change in domestic surplus. When we include the benefits of exporters, however, total surplus increases by 5.37 percent. The story in Germany is a bit different. Consumer gains outweigh domestic firm losses in Germany and domestic surplus increases by 4.96 percent. Essentially, removing border frictions improves German site owners access to high-productivity Danish firms and erodes Enercon's substantial market power in Germany. When we include the benefits to Danish exporters, elimination of the border raises surplus in the German market by a substantial 10.42 percent.

We conclude this section by repeating an important disclaimer. Our second counterfactual represents an elimination of all border frictions. In reality, these frictions are generated by a complex combination of political, administrative, and cultural differences between countries. It is unlikely that any policy initiative would succeed in eliminating these differences completely. Rather, our findings illustrate the magnitude of the border and its effect on firms and consumers in the wind turbine industry. Policy makers may view the results as an upper bound on what can be accomplished through political integration.

6 Reduced-Form Estimation of the Structural Border

Several studies have used a no-arbitrage condition to motivate estimates of border frictions. In contrast, we have explicitly modeled border costs within an oligopolistic framework, without appealing to the law of one

[36] Of course, these declines do not account for benefits realized in the export market. See Table 5 for an accounting of how each firm fairs as both an domestic producer and an exporter under our counterfactual scenarios.

price directly. In order to highlight how these approaches may differ, this section uses our model-generated prices in a reduced-form regression that relates price differentials across space to distance and the border. Our goal is to compare the implied width of the border from this exercise with the structural estimate on variable border costs from Section 4.1 and identify the sources of discrepancies between the two.[37]

The regression is in the spirit of Engel and Rogers (1996) and the ensuing literature that estimate the border effect using within- and cross-country prices. In this line of inquiry, the border effect is the additional price dispersion brought about by national boundaries. The researcher starts with a panel data p^j_{kt} of prices for identical, tradable goods indexed by j measured in locations indexed by k. In order to test deviations from the relative version of the law of one price, she collapses the time series variation to cross-sectional variation in the volatility of prices across goods and locations. To be specific, let $\sigma^j_{k\ell}$ represent the standard deviation of period-to-period changes in the relative price over time of good j in locations k and ℓ.

$$\sigma^j_{k\ell} \;=\; \text{std}\left(\left\{\frac{p^j_{kt}}{p^j_{\ell t}} - \frac{p^j_{kt-1}}{p^j_{\ell t-1}}\right\}_{t=1}^{T}\right)$$

A low $\sigma^j_{k\ell}$ means that shocks to the price of good j in one location are quickly transmitted to the other. Thus, the higher the volatility, the weaker is the law-of-one price between cities k and ℓ. A typical regression à la Engel and Rogers (1996) is as follows:

$$\sigma^j_{k\ell} \;=\; \delta^j_d \cdot \ln \text{distance}_{k\ell} + \delta^j_b \cdot \text{border}_{k\ell} + \delta^j_k + \delta^j_\ell + \epsilon^j_{k\ell},$$

where $\text{distance}_{k\ell}$ is the distance between locations k and ℓ, and $\text{border}_{k\ell} = 1$ if k and ℓ are in the different countries. Location fixed effects are included to control for city-specific differences that impact price volatility, such as different seasonal patterns or data collection techniques. The border effect for good j is then interpreted in terms of the distance equivalent of the border dummy coefficient:

$$\textit{Border Effect}_j \;=\; \exp(\delta^j_b/\delta^j_d).$$

Rather than consider volatility, our framework allows us to estimate the border effect from the absolute version of the law of one price using price differentials directly.[38] In our thought experiment, we treat the border width of 432 kilometers estimated in Section 4.1 as its "true" value. An econometrician trying to recover it from a statistical model observes prices p^j_k for the same turbine j in different locations k, the

[37] We compare this estimate to the variable border cost because the fixed border cost is sunk when firms set prices.

[38] Engel and Rogers (1996) do not test the absolute law of one price directly because of two reasons. First, measured prices are typically indices rather than actual prices. Second, price differentials can arise due to differences in local market conditions and input costs that are not traded.

Table 7: Border Effect Estimates from Reduced-Form Estimation

Firm	$\hat{\delta}_b/\hat{\delta}_d$ in km
Bonus	741.98
Nordtank	857.95
Micon	781.16
Vestas	516.61
WindWorld	1092.08
Nordex	3472.55

distances between locations ($\text{distance}_{k\ell}$) but not the distances between locations and producers (distance_{kj}). This is a good description of the information set used by researchers who estimate reduced-form regressions depicted above. We follow their practice and estimate a similar OLS regression, adapted to our framework:

$$|p_k^j - p_\ell^j| \;=\; \delta_d^j \cdot \text{distance}_{k\ell} + \delta_b^j \cdot \text{border}_{k\ell} + \delta_k^j + \delta_\ell^j + \epsilon_{k\ell}^j. \tag{11}$$

Our data allows us to calculate distances between each project pair (k, ℓ). Again, the border dummy equals one if two projects are in different countries. Price p_k^j is the endogenous equilibrium price bid of firm j for project k given in (3) in Section 3.1. We report the implied border effect (δ_b^j/δ_d^j) from this regression in Table 7. Evidently, this exercise overestimates the border effect in our model for all producers.[39] For Danish firms, estimates vary between 1.2 to 2.5 times the "true" value of 432 kilometers. The bias is much higher for the German firm, Nordex.

To gain intuition on the sources of this overestimation, contrast (11) with the price difference implied by our structural model using our estimates $(\hat{\beta}_d, \hat{\beta}_b, \hat{\rho}_{kj})$ in expressions (2) and (3):

$$|p_k^j - p_\ell^j| \;=\; \Big|\hat{\beta}_d\big(\text{distance}_{kj} - \text{distance}_{\ell j}\big) + \hat{\beta}_b\big(\text{border}_{kj} - \text{border}_{\ell j}\big) + \Big(\frac{1}{1-\hat{\rho}_{kj}} - \frac{1}{1-\hat{\rho}_{\ell j}}\Big)\Big| \tag{12}$$

The three terms in this equation reflect the sources of producer-level spatial price differentials in our model: differences in project-to-producer distances are captured by the first term, differences in border frictions for each project are captured by the second term, and differences in project-specific mark-ups due to variation in competitive structure across space are captured by the last term. Note that the firm competitiveness parameter has canceled out through taking differences.

[39] The overestimation is robust to whether or not we include location fixed effects, which are included in the reported results. In the underlying regressions, distance and border coefficients are statistically significant at .01 level for all producers. The detailed regression results are available from authors upon request.

When we compare this data generating process with equation (11), it is apparent that the linear reduced-form regression is misspecified. In the structural equation (12), price differentials are generated by the absolute value of several *differences* in project-to-producer distances, destination countries, and mark-ups, whereas (11) is a linear function of related, but different, variables. While trying to emulate this model-based expression, equation (11) suffers from two additional problems: First, using project-to-project distances ($\text{distance}_{k\ell}$) instead of the differences in project-to-producer distances differences ($\text{distance}_{kj} - \text{distance}_{\ell j}$) leads to (non-classical) measurement error. The triangle inequality implies that the actual difference of the project-to-producer distances is less than the project-to-project distances. This would tend to bias the estimate of δ_d towards zero relative to the distance parameter $\hat{\beta}_d$ in (12).[40] Second, (11) suffers from omitted variable bias due to the mark-up differentials being left out. Note that the vector of location fixed effects included in the regression cannot properly characterize the mark-up differences between project-pairs, since those dummies would capture information about levels instead of differences. Since markup differences are likely to be correlated with the border dummy, this would tend to bias δ_b upwards due to an endogeneity problem. The combined result is the border effect estimates $\hat{\delta}_b/\hat{\delta}_d$ in Table 7 are higher than their structural analogue, $\hat{\beta}_b/\hat{\beta}_d$.

While our thought experiment focuses on price deviations directly rather than price volatility, it is easy to see that the linear specification error, measurement error and omitted variable bias would arise when volatility measures are the dependent variable. The findings of this section resonate with Gorodnichenko and Tesar (2009) who argue that model-free border-effect regressions fail to identify border frictions when there is within-country price dispersion due to spatial variation in competition and transportation costs. Moreover, we show the importance of using disaggregated data—in our case the knowledge of manufacturing locations—to properly control for variation in distance costs and markups. These issues apply to a large range of industries in which specific producers operate in only a few locations and the set of active firms is different on either side of the border.

7 Conclusion

This paper uses transaction-level data for a specific industry to document the impact of fixed and variable border costs while controlling for several sources of bias that plague analysis of aggregated trade flows. The model and the detailed geographical information on manufacturers and projects allow us to better

[40] The triangle inequality discrepancy explains why the measured effect is so much higher for Nordex in Table 7. Danish firms are all located at the north end of the set of projects. Hence, project-to-project distance is a better proxy for the distance differential, since the majority of projects are south of their manufacturing facility. Nordex, however, is more centrally located. As a result, two separate projects in Denmark and Germany that are equidistant to Nordex, and thus have a low firm-to-project distance differential will have a high project-to-project distance. Nordex's distinctively higher border effect estimate is thus in part due to a poorer distance proxy for many project pairs.

control for distance costs and spatial differences in competition on either side of the border than the existing literature. The model combines conventional tools from the literature into a novel approach to analyze spatial oligopolistic competition in a multi-country setting.

The large differences in national market shares in the wind turbine industry between Denmark and Germany arise not only through costs associated with distance, but also through barriers to foreign market entry and higher variable costs associated with crossing the border. These border costs are substantial; more than 50 percent of gap in cross-border market shares can be attributed to them. Our results also indicate that the welfare gains from a hypothetical removal of all border frictions between Germany and Denmark—including barriers that are difficult to remove, such as language—are substantial. We cannot, however, separately identify the roles that bureaucratic, linguistic, or cultural differences play in generating border frictions.

Nonetheless, the existence of large border frictions within the European wind turbine industry has important policy implications for the EU. Due to growing concerns about climate change, many governments, including EU members and the United States, subsidize renewable energy generation. The efficiency of subsidies in the wind electricity output market is closely related to the degree of competition in the upstream market for wind turbines themselves. If there are substantial frictions to international trade in turbines, a national subsidy to the downstream market may implicitly be a subsidy to domestic turbine manufacturers. This is against the intensions of EU common market policy, which seeks to prevent distortions due to subsidies given by member states exclusively to domestic firms. In fact, Denmark, which has one of the most generous wind energy subsidies in Europe, is also home to the most successful European producers of wind turbines. Given our findings of large border frictions in the upstream market, EU members may wish to harmonize renewable energy tariffs to ensure equal treatment of European firms in accordance with the principles of the European single market project.

References

Anderson, J. E., and E. van Wincoop (2003): "Gravity with Gravitas: A Solution to the Border Puzzle," *The American Economic Review*, 93(1), pp. 170–192.

Anderson, P. (2007): "Scout Moor Wind Farm Under Construction," `http://www.geograph.org.uk/gallery/scout_moor_wind_farm_under_construction_6568`.

Balistreri, E. J., and R. H. Hillberry (2007): "Structural estimation and the border puzzle," *Journal of International Economics*, 72(2), 451–463.

Benkard, C. L. (2004): "A Dynamic Analysis of the Market for Wide-Bodied Commericial Aircraft," *Review of Economic Studies*, 71, 581–611.

Besanko, D., U. Doraszelski, Y. Kryukov, and M. Satterthwaite (2010): "Learning-by-Doing, Organizational Forgetting, and Industry Dynamics," *Econometrica*, 78(2), 453–508.

Bresnahan, T. F., and P. Reiss (1991): "Empirical Models of Discrete Games," *Journal of Econometrics*, 48(1), 57–81.

Broda, C., and D. E. Weinstein (2008): "Understanding International Price Differences Using Barcode Data," NBER Working Papers 14017, National Bureau of Economic Research, Inc.

Caplin, A., and B. Nalebuff (1991): "Aggregation and Imperfect Competition: On the Existence of Equilibrium," *Econometrica*, 59(1), 25–59.

Ciliberto, F., and E. Tamer (2009): "Market Structure and Multiple Equilibria in Airline Markets," *Econometrica*, 77(6), 1791–1828.

Das, S., M. J. Roberts, and J. R. Tybout (2007): "Market Entry Costs, Producer Heterogeneity and Export Dynamics," *Econometrica*, 75(3), 837–873.

Eaton, J., and S. Kortum (2002): "Technology, Geography, and Trade," *Econometrica*, 70(5), pp. 1741–1779.

Eaton, J., S. Kortum, and F. Kramarz (2011): "An Anatomy of International Trade: Evidence From French Firms," *Econometrica*, 79(5), 1453–1498.

Engel, C., and J. H. Rogers (1996): "How Wide Is the Border?," *The American Economic Review*, 86(5), pp. 1112–1125.

Ericson, R., and A. Pakes (1995): "Markov-Perfect Industry Dynamics: A Framework for Empirical Work," *Review of Economic Studies*, 62(1), 53–82.

Fowlie, M., M. Reguant, and S. P. Ryan (2011): "Market-based Emissions Regulation and the Evolution of Market Structure," MIT Mimeo.

Franken, M., and T. Weber (2008): "Heavy Duty," *New Energy*, 2008(5), 28–37.

Goldberg, P. K., and F. Verboven (2001): "The Evolution of Price Dispersion in the European Car Market," *The Review of Economic Studies*, 68(4), 811–848.

Goldberg, P. K., and F. Verboven (2005): "Market integration and convergence to the Law of One Price: evidence from the European car market," *Journal of International Economics*, 65(1), 49 – 73.

Gopinath, G., P.-O. Gourinchas, C.-T. Hsieh, and N. Li (2011): "International Prices, Costs and Markup Differences," *American Economic Review*, 101(6), 2450–86.

Gorodnichenko, Y., and L. L. Tesar (2009): "Border Effect or Country Effect? Seattle May Not Be So Far from Vancouver After All," *American Economic Journal: Macroeconomics*, 1(1), 219–41.

Hillberry, R., and D. Hummels (2008): "Trade responses to geographic frictions: A decomposition using micro-data," *European Economic Review*, 52(3), 527 – 550.

Hillberry, R. H. (2002): "Aggregation bias, compositional change, and the border effect," *Canadian Journal of Economics*, 35(3), 517–530.

Holmes, T. J. (2011): "The Diffusion of Wal-Mart and the Economies of Density," *Econometrica*, 79(1), 253–302.

Holmes, T. J., and J. J. Stevens (2012): "Exports, Borders, Distance, and Plant Size," *Journal of International Economics, forthcoming*.

Interessenverband Windkraft Binnenland (various years): "Windkraftanlagen Marktübersicht," 1994 through 1996.

Judd, K. L., and C.-L. Su (2011): "Constrained Optimization Approaches to Estimation of Structural Models," *Econometrica, forthcoming*.

McCallum, J. (1995): "National Borders Matter: Canada-U.S. Regional Trade Patterns," *The American Economic Review*, 85(3), 615–623.

Melitz, M. J. (2003): "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity," *Econometrica*, 71(6), 1695–1725.

Morales, E., G. Sheu, and A. Zahler (2011): "Gravity and Extended Gravity: Estimating a Structural Model of Export Entry," Harvard University Mimeo.

Obstfeld, M., and K. Rogoff (2001): "The Six Major Puzzles in International Macroeconomics: Is There a Common Cause?," in *NBER Macroeconomics Annual 2000, Volume 15*, NBER Chapters, pp. 339–412. National Bureau of Economic Research, Inc.

Pakes, A., J. Porter, K. Ho, and J. Ishii (2006): "Moment Inequalities and their Application," Harvard University and University of Wisconsin.

Pinkse, J., and M. E. Slade (1998): "Contracting in space: An application of spatial statistics to discrete-choice models," *Journal of Econometrics*, 85(1), 125 – 154.

Roberts, M. J., and J. R. Tybout (1997): "The Decision to Export in Colombia: An Empirical Model of Entry with Sunk Costs," *American Economic Review*, 87(4), 545–64.

Romalis, J. (2007): "Capital Taxes, Trade Costs, and The Irish Miracle," *Journal of the European Economic Association*, 5(2-3), 459–469.

Sijm, J. (2002): "The Performance of Feed-in Tariffs to Promote Renewable Electricity in European Countries," Energy Resource Center of the Netherlands, ECN-C–02-083.

Appendices

Appendix A Data

A.1 Description

The register of Danish wind turbines is publicly available from the Danish Energy Agency (http://www.ens.dk/en-US/Info/FactsAndFigures/Energy_statistics_and_indicators/OverviewOfTheEnergySector/RegisterOfWindTurbines/Sider/Forside.aspx). This dataset spans the entire universe of Danish turbine installations since 1977 until the most recent month. The data on German installations is purchased from the private consulting company Betreiber-Datenbasis (http://www.btrdb.de/). Typically, several turbines are part of one wind farm project. The German data comes with project identifiers. We aggregate Danish turbines into projects using the information on installation dates, cadastral and local authority numbers. Specifically, turbines installed in the same year, by the same manufacturer, under the same cadastral and local authority number are assigned to the same project. The fine level of disaggregation provided by cadastral and local authority numbers minimize the measurement error.

Data on manufacturer locations was hand-collected from firms' websites and contacts in the industry. As of 1995 and 1996, seven out of ten large firms we use for our analysis were operating a single plant. Bonus, Vestas and Nordex had secondary production facilities. For these firms, we use the headquarters. Our industry contacts verified that these headquarters were also primary production locations with the majority of value-added. Equipped with the coordinates of projects and production locations, we calculated road distances as of June 2011 using the Google Maps API (http://code.google.com/apis/maps/). Therefore, our road distances reflect the most recent road network. For developed countries such as Germany and Denmark, we believe the error introduced by the change in road networks over time is negligible. Using direct great-circle distances in estimation generated virtually the same results.

A.2 Project Characteristics

Table 8, and Figures 8-10 provide some summary statistics on project characteristics in the two countries. Differences in distance to producers reflect heterogeneity in country size. Evidently, key observable characteristics such as electricity generating capacity, tower height and rotor diameter are remarkably similar in the two markets, ruling out product differentiation as a source of market segmentation. Slightly higher tower heights in Germany are due to lower wind speeds in southern regions. In such an environment, larger turbines are needed to attain the same capacity. What matters for this paper is that wind conditions do not change at the border. The European wind atlas available at the following link verifies that this is the case. (http://www.wind-energy-the-facts.org/en/appendix/appendix-a.html).

A.3 List Prices

The survey of the German wind turbine market published by Interessenverband Windkraft Binnenland (various years) provides information on list prices for various turbine models as advertised by producers. These prices, however, are only suggestive and do not reflect project-specific final transaction prices. We use this information to verify the validity of our CRTS assumption. Figure 11 plots the per kilowatt price of various models against their total kilowatt capacity. Evidently, there are increasing returns up to 200 KWs. Beyond that range, per unit prices are mostly flat. As Figure 10 shows, a majority of the turbines installed in this period were in the 400-600 KW range.

A.4 Regression Discontinuity Design

We estimate the following local linear probability model in Subsection 2.2 to implement the regression discontinuity design:

Table 8: Summary Statistics of Projects

		Denmark	Germany
Capacity (KW)	Mean	475.81	472.59
	St. Dev.	207.93	175.98
	Median	600	500
	$10th$ percentile	225	225
	$90th$ percentile	600	600
Tower height (m)	Mean	38.34	49
	St. Dev.	7.96	8.64
	Median	40	50
	$10th$ percentile	30	40
	$90th$ percentile	46	65
Rotor diameter (m)	Mean	37.43	38.51
	St. Dev.	9.13	7.02
	Median	42	40.3
	$10th$ percentile	29	29.5
	$90th$ percentile	44	44
Distance to the border (km)	Mean	159.38	296.88
	St. Dev.	72.33	162.23
	Median	169.45	295.12
	$10th$ percentile	51.59	90.68
	$90th$ percentile	242.58	509.20
Distance to producers* (km)	Mean	154.02	366.58
	St. Dev.	31.26	100.19
	Median	169.45	344
	$10th$ percentile	117.52	258.20
	$90th$ percentile	192.65	510.78
Number of turbines per project	Mean	1.94	1.95
	St. Dev.	2.07	2.52
Number of projects	1995-1996	296	930
	1997-2005	1373	4148

Notes: Summary statistics of product characteristics in the first six panels are from the sub-sample of projects installed in 1995-1996. Onshore projects only.
(*): Average distance to firms with positive sales in that market.

Table 9: RDD Results

	(1) 1995-1996	(2) 1997-1998	(3) 1999-2005
Germany	-0.295*	-0.253*	-0.318**
	(-2.13)	(-2.24)	(-2.67)
Constant	0.922***	0.911***	0.862***
	(7.56)	(11.45)	(8.00)
Observations	1189	1237	3318
Adjusted R^2	0.289	0.380	0.192

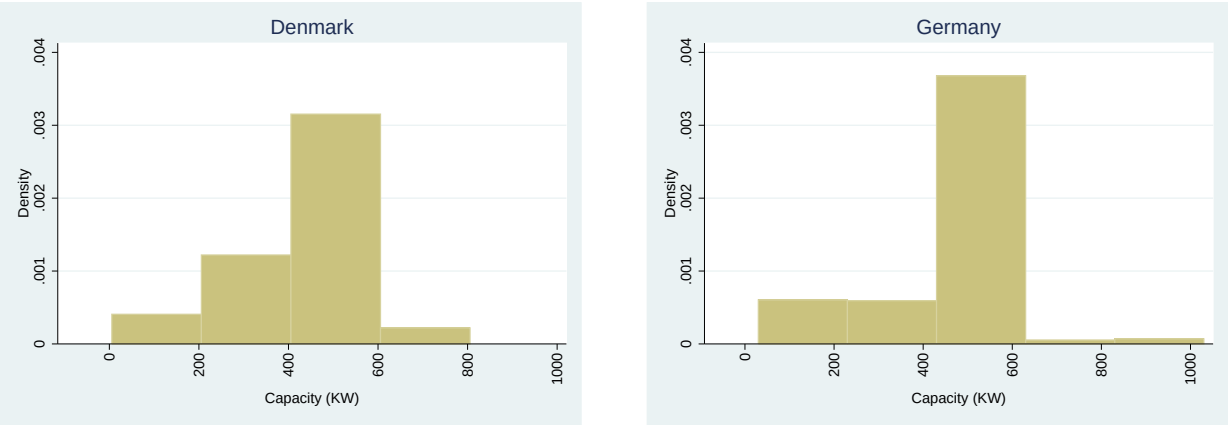
Notes: t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

$$y_i = \alpha_0 + \sum_{k=1}^{k=3} \alpha_k \cdot \text{distance}_i^k + \gamma \cdot \text{Germany}_i + \sum_{k=1}^{k=3} \eta_k \cdot \text{distance}_i^k \cdot \text{Germany}_i + \epsilon_i. \quad (13)$$

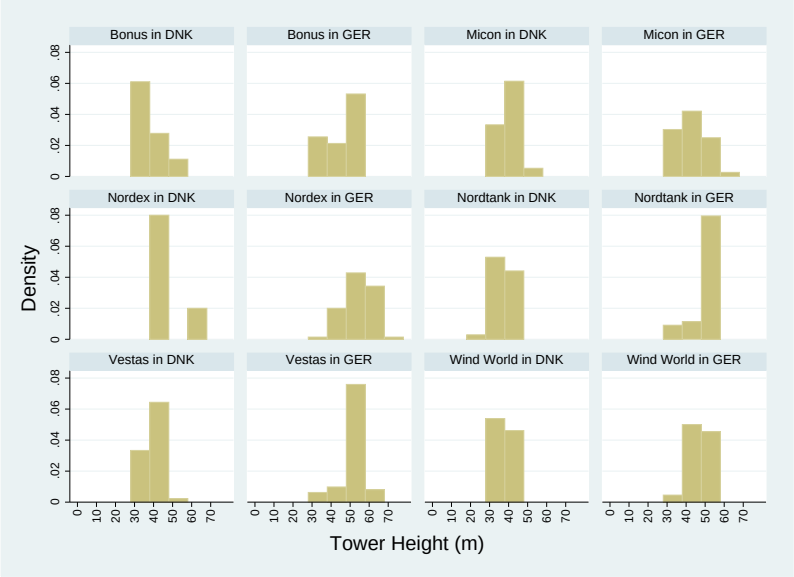
The dependent variable is $y_i = 1$ if the producer of project i is one of the five large Danish firms in our model (Bonus, Micon, Nordtank, WindWorld or Vestas), and zero otherwise. distance_i is the distance to the border. The effect of the border is picked up by the dummy variable Germany_i that equals one if the project is in Germany, and zero otherwise. The parameter of interest is γ. The band we use for distance is [-300km, 600km]. We run the estimation for three subperiods: 1995-1996, 1997-1998 and 1999-2005. The last subperiod pools data over seven years to ensure that there are enough observations in the neighborhood of the border at both sides. This becomes an issue because of the saturation of the Danish market after late 1990s. Table 9 reports the results for significant variables. In all cases, Germany dummy is negative and statistically significant at the 5% level. Moreover, the border effect is very stable over time. This verifies that we are not focusing on a peculiar period by using data from 1995-1996 in our structural estimation. Figure 5 is the fit of the RDD estimation.

Figure 8: KW Capacity Histograms by Market



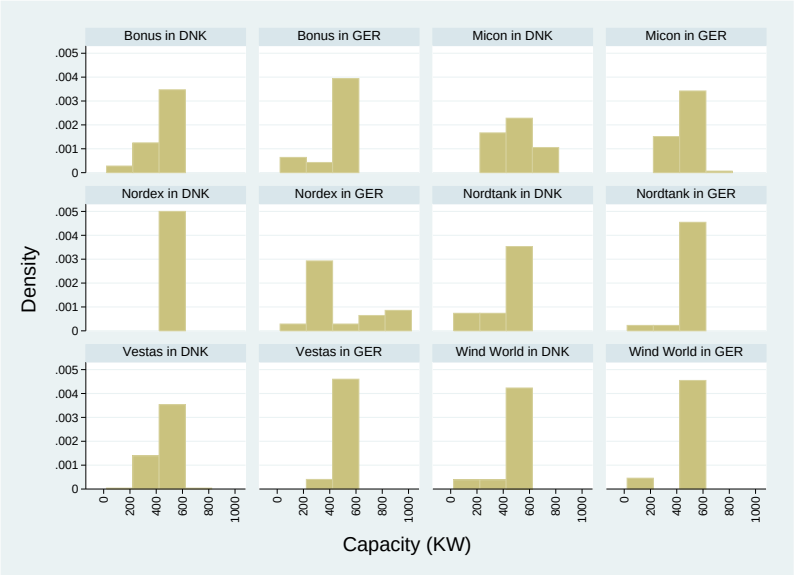
Notes: An observation is average kw capacity of turbines in a project. Years 1995 and 1996 only.

Figure 9: Tower Height Histograms by Producer and Market



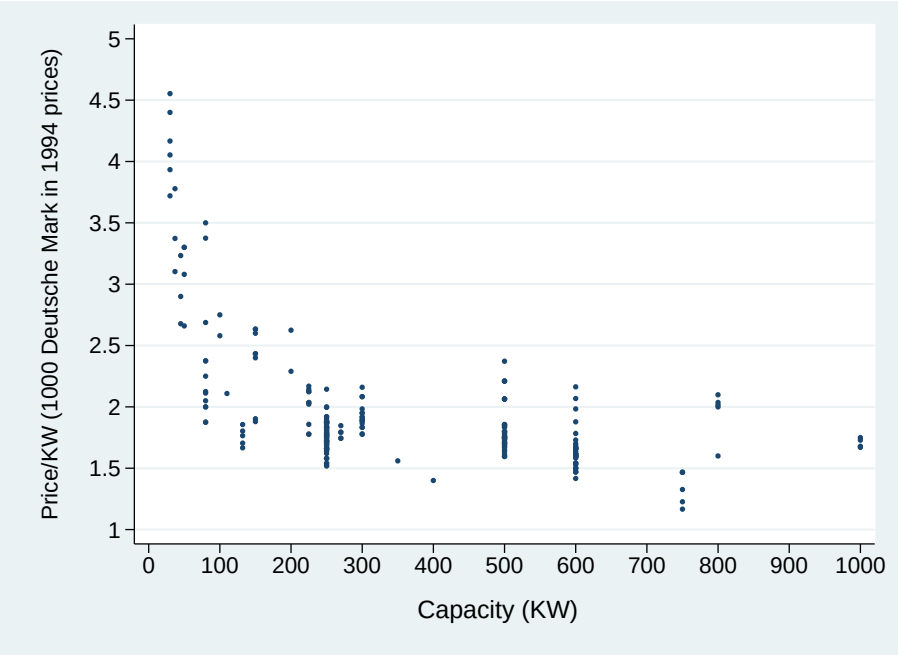
Notes: An observation is average tower height of turbines in a project. Years 1995 and 1996 only. "Bonus in DNK (GER)" indicates projects supplied by Bonus in Denmark (Germany).

Figure 10: KW Capacity Histograms by Producer and Market



Notes: An observation is average kw capacity of turbines in a project. Years 1995 and 1996 only. "Bonus in DNK (GER)" indicates projects supplied by Bonus in Denmark (Germany).

Figure 11: Per KW List Prices of Various Turbines Offered in 1995-1996



Notes: Pooled over all producers.

Appendix B Robustness to Local Unobservables and Economies of Density

In order to derive the pricing equation, our model assumes that the unobservable shock to managers profits, ϵ_{ij}^{ℓ}, is unknown to firms, but drawn from a known distribution which is independent across projects and firms. Thus, we abstract away from the existence of spatial autocorrelation of unobservables across projects. This section assesses whether this assumption has the potential to bias our estimate of the border effect.

There are several reasons for being concerned about the independence assumption. The assumption will be violated if firms directly observe sources of firm-project cost variation which are not explicitly controlled for by the model. While we feel that firms' productivity levels, firm-project distances, and the border dummy are the primary determinants of costs, other potential sources of variation could relate to unobservable local conditions being more amenable to a particular firm (e.g., local politics or geographic features of an area could result in lower cost for some firms). The independence assumption will also be violated if economies of density can be realized by a firm constructing several projects located geographically close together. Economies of density might be present if, for example, clustering projects together reduces travel costs for routine maintenance. Such economies of density might make the individual projects less expensive to maintain on a per-unit basis, leading firms with nearby installed projects to have a cost advantage over other firms that is not recognized in our model.

The existence of local unobservables would generate spatial autocorrelation in the error terms between projects which are geographically close. This would violate our assumption that the errors are independent across projects. Moreover, if firms are responding to economies of density of projects, firms pricing decisions become dynamic in nature. Since winning a project today lowers the firms' costs on other projects in the future, firms would not choose prices to maximize project-level profits, but rather the present discounted value of profits on this project and future projects. Both of these forces would lead firm's projects to be more tightly clustered together than our model would predict, leading to spatial autocorrelation in firms' error terms across projects. To test for the presence of spatial autocorrelation, we consider the following parametric model for the error term,

$$\epsilon_j = \gamma + \psi W \epsilon_j + \nu_i. \tag{14}$$

Here, ϵ_j is the vector of private shocks for firm j in all projects, γ is Euler's constant—the mean of the extreme value distribution, W is a known spatial weight matrix that determines the degree of influence one project has on another, and ν_i are independent and identically distributed mean-zero shocks. The scalar ψ determines the degree of spatial autocorrelation, we wish to test the null hypothesis that spatial autocorrelation is not present, i.e., that $\psi = 0$ and the ϵ_{ij} are in fact independent across projects.

In order to perform the test, we must specify the spatial weight matrix W. An element of the spatial weight matrix, W_{ik} provides an indication of how strongly project k is related to project i. Clearly many different specifications are possible, including inverse distance (measured either directly or though a road network), inclusion within the same region, or nearest neighbor adjacency. In practice we specify W as,

$$W_{ik} = \begin{cases} 1 & \text{if } dist(i,k) < 30\text{km}, \\ 0 & \text{otherwise}, \end{cases}$$

where distance is the direct distance (as the crow flies) in kilometers between projects i and j.[41]

We are unable to directly test for spatial autocorrelation in ϵ_{ij}^{ℓ} because as with all discrete choice models, ϵ_{ij}^{ℓ} is not directly recoverable. Instead, we follow Pinkse and Slade (1998) and test our results for spatial autocorrelation using the generalized errors. The generalized errors are the expectation of ϵ_{ij}^{ℓ} conditioned on the observed data and the model being correctly specified. Given the structure of our model, the generalized errors can be derived using the extreme-value density function,[42]

$$\hat{\epsilon}_{ij}^{\ell} = \begin{cases} \gamma - \log \rho_{ij}^{\ell} & \text{if } y_{ij}^{\ell} = 1, \\ \gamma + \frac{\rho_{ij}^{\ell}}{1-\rho_{ij}^{\ell}} \log \rho_{ij}^{\ell} & \text{if } y_{ij}^{\ell} = 0. \end{cases}$$

Again γ, represents Euler's constant—the unconditional expectation of the extreme value distribution. While the derivation of these expectations is algebraically tedious, the result is intuitive: the more likely a manufacturer j is to be selected by the project manager, the lower ϵ_{ij}^{ℓ} must be in order for selection to occur. Hence, $\hat{\epsilon}_{ij}^{\ell}$ is decreasing in the ex-ante probability of firm j being selected. The fact that the distribution of $\hat{\epsilon}_{ij}^{\ell}$ conditional on j not being chosen is independent of the actual choice observed in market i is a consequence of the well known independence of irrelevant alternatives (IIA) property of extreme-value discrete choice models. Note that, if the null hypothesis of no auto-correlation is violated, $\hat{\epsilon}_{ij}^{\ell}$ will be mis-specified. Nonetheless, they are useful to conduct a hypothesis test that $\psi = 0$.

We can use ordinary least squares to estimate ψ from the equation,

$$\hat{\epsilon}_j = \gamma + \psi W \hat{\epsilon}_j + \nu_i$$

and test whether $\psi = 0$. Note that, the estimate we generate, $\hat{\psi}$, is only consistent under the null hypothesis since the null is assumed in the construction of $\hat{\epsilon}_j$ and ordinary least squares is only consistent if $\psi = 0$.

The results are reported in Table 10.[43] While the magnitude of the estimated $\hat{\psi}$ is small, the test strongly rejects the null hypothesis for every firm, due in part to the the high precision of the estimates. We conclude that some degree of spatial autocorrelation is present, although it appears to be mild.

The presence of spatial autocorrelation has the potential to bias our estimate of the border effect. In particular, if spatial autocorrelation is due to cost or demand advantages in installing near already completed projects constructed by the same manufacturer, and if exporters have a smaller installed base within a country than do domestic firms, then the border effect may be capturing differences in the installed bases of foreign and domestic firms in addition to the variable cost of exporting. Alternatively, if serial correlation is due to local unobserved characteristics then the location of previous installations, while not cost reducing in and

[41] Our results are robust to raising or lowering the distance cutoff and using a specification of W based on inverse distance.

[42] The derivation is available from the authors upon request.

[43] It is important that the test be conducted with heteroskedasticity-robust variance estimates, since there is little reason to believe that the generalized errors are homoscedastic.

Table 10: Results from Auto-Correlation Tests

Manufacturer	$\hat{\psi}$	Std. Error	t-Stat.
Fringe	0.024	0.008	3.096
Bonus (DK)	0.027	0.005	5.079
Nordtank (DK)	0.024	0.004	6.122
Micon (DK)	0.032	0.004	7.225
Vestas (DK)	0.034	0.005	7.124
WindWorld (DK)	0.031	0.007	4.635
Enercon (DE)	0.043	0.007	6.000
Fuhrlaender (DE)	0.034	0.005	7.165
Nordex (DE)	0.048	0.006	8.194
Suedwind (DE)	0.038	0.014	2.757
Tacke (DE)	0.029	0.005	6.118

of themselves, serve as proxies for unobservable local conditions. In this spirit, we propose the following specification to check the robustness of our results to mild spatial autocorrelation. We re-estimate the model with the augmented cost function,

$$c_{ij} = \phi_j + \beta_d \cdot \text{distance}_{ij} + \beta_b \cdot \text{border}_{ij} + \beta_c \cdot \text{installed}_{ij},$$

where,[44]

$$\text{installed}_{ij} = \begin{cases} 1 & \text{if firm } j \text{ installed a turbine within 30km of project } i \text{ between 1991 and 1994,} \\ 0 & \text{otherwise.} \end{cases}$$

The new coefficient, β_c is able to capture the relationship between previously installed turbines and the costs of future projects. We are unable, however, to determine whether β_c is a causal effect, a proxy for local unobservables, or some combination of the two. Firms within our model continue to price according to static profit maximization. They do not take into account the possibility that building a turbine will make nearby projects less expensive in the future. This is consistent with the idea that the existence of local installations being merely a proxy variable and having no causal impact on future costs.

The results from this robustness specification are presented in Table 11. The coefficient on having a nearby installation has the expected negative sign (nearby installations are indicative of lower costs) and is of substantial magnitude, but is statistically insignificant. The estimates of both distance costs, β_d and variable border costs, β_b both decrease slightly, but remain strongly significant. The estimated impact of the border relative to distance actually increases from 432 km to 502 km. Overall, these results appear to indicate that while unobservable local conditions of economies of density may induce some spatial autocorrelation between projects, the effect is mild and is not substantially impacting our primary results on the size of the border effect. In future work, we hope to investigate whether there is a causal effect of installations on the cost of future projects, but this question will require a fully dynamic pricing model which is outside the scope of our investigation of border costs.

[44] We have also experimented with a using distance to the nearest installed project in the cost function and using only projects installed between 1993 and 1994, and have found qualitatively similar results.

Table 11: Robustness Check: Nearby Installed Turbines

	Coefficient	Std. Error
Border Variable Cost, β_b	0.688	(0.178)
Distance Cost (100km), β_d	0.137	(0.031)
Nearby Installation, β_c	-1.249	(1.167)
Firm Fixed Effects, ξ_j		
Bonus (DK)	1.256	(0.189)
Nordtank (DK)	1.462	(0.183)
Micon (DK)	2.046	(0.160)
Vestas (DK)	2.689	(0.170)
WindWorld (DK)	0.640	(0.211)
Enercon (DE)	2.719	(0.147)
Fuhrlaender (DE)	-0.010	(0.266)
Nordex (DE)	0.858	(0.184)
Suedwind (DE)	-0.187	(0.206)
Tacke (DE)	1.578	(0.152)
Log-Likelihood	-2286.15	
N	1226	

Appendix C Computational Method

C.1 Estimation of the Project Bidding Game

We formulate the estimation of the project bidding game as a constrained optimization problem.[45] The objective is to maximize the likelihood function subject to satisfying the firm-project specific winning probabilities expressions that come out of our model. We reformulate the problem defined in (10) for the computational implementation. The reformulated constraints are mathematically equivalent to those in (10). They come with two major advantages: First, when we reformulate the system maximizing the log-likelihood instead of the likelihood function, and rewrite the constraints, we are removing most of the nonlinearity. Second, winning probabilities only affect their respective equation and the adding-up constraint for the respective project. The sparse structure of the Jacobian of the constraints makes this large optimization problem feasible. The reformulated problem is

$$\begin{aligned}
\max_{\theta,\,\rho} \quad & \sum_{\ell\in\{D,G\}}\sum_{i=1}^{N_\ell}\sum_{j=1}^{|\mathcal{J}_\ell|} y_{ij}^{\ell}\log\rho_{ij}^{\ell} \\
\text{subject to:} \quad & \log\rho_{ij}^{\ell}-\log\rho_{i0}^{\ell}=\xi_j-\beta_d\cdot\text{distance}_{ij}-\beta_b\cdot\text{border}_{ij}^{\ell}-\frac{1}{1-\rho_{ij}^{\ell}} \\
& \sum_{k=1}^{|\mathcal{J}_\ell|}\rho_{ik}^{\ell}+\rho_{i0}^{\ell}=1 \\
& \text{for } \ell\in\{D,G\},\, i\in\{1,...,N_\ell\},\, j\in\mathcal{J}.
\end{aligned}$$

For the baseline estimation, there are 11 constraints for every German project, and 7 constraints for every Danish project ($|\mathcal{J}_G|=10$ and $|\mathcal{J}_D|=6$ plus one fringe firm in every market). Since we have 930

[45] See Judd and Su (2011) for a seminal paper that explains why constrained optimization of structural models is often superior to estimation via nested fixed points.

German and 296 Danish project this aggregates to 12,302 constraints. In our baseline specification we are choosing 12,314 variables (12 structural parameters and 12,302 equilibrium win probabilities for each firm in each market)

We use the constrained optimization solver KNITRO to solve the problem. To improve speed and accuracy of the estimation, we hand-code the analytical derivatives of the object of function and the constraints and provide the sparsity structure of the Jacobian to the solver.[46] In order to find a global maximum we pick 10 random starting values for the structural parameters. The estimation converges to the same solution for all attempted starting values.

We calculate the covariance matrix of the parameter estimates using the outer product rule.

1. First, we calculate the score of each winning firm project pair, $\partial \log \rho_i^* / \partial \theta$, using numerical derivatives. This involves perturbing the $\widehat{\theta}$ vector. Note that the step size to perturb θ should be larger than the numerical tolerance level of the equilibrium constraints. Then the equilibrium constraints are resolved.

2. We then calculate the inverse of the covariance matrix:

$$\widehat{S}(\widehat{\theta}) = \sum_{i=1}^{N} \frac{\partial \log \rho_i^*(\widehat{\theta})}{\partial \theta} \frac{\partial \log \rho_i^*(\widehat{\theta})}{\partial \theta}'$$

C.2 Counterfactuals

The point estimate $\widehat{\theta}$ automatically satisfies the equilibrium constraints in the benchmark scenario with fixed entry and variable border costs. In the counterfactual "No fixed border costs" we use $\widehat{\theta}$ to then resolve the equilibrium constraints, with every firm being active in every market, $|\mathcal{J}_D| = |\mathcal{J}_G| = 10$. In the counterfactual "No border costs" we resolve the same system of equilibrium constraints with the variable border cost coefficient set to zero.

We use a parametric bootstrap procedure to calculate the standard errors for our counterfactuals. We draw 200 parameter vectors from the distribution of estimated parameters (multivariate normal distribution with mean θ and covariance matrix $\widehat{S}(\widehat{\theta})^{-1}$). First we resolve the baseline equilibrium constraints, then the constraints for the scenario with no fixed entry costs, and finally the constraints for the no border costs scenario (with each firm active in every market and the variable border costs coefficient set to zero). We store the equilibrium outcomes from each of these draws and use them to calculate the standard errors for our counterfactuals.

[46] Prior to the estimation we check via finite differences that our analytical gradients are correct.

Measuring the Competitiveness Benefits of a Transmission Investment Policy: The Case of the Alberta Electricity Market

by

Frank A. Wolak*
Director, Program on Energy and Sustainable Development
Professor, Department of Economics
Stanford University
Stanford, CA 94305-6072
wolak@zia.stanford.edu

July 2, 2012
Preliminary Draft

*This research was supported by the Alberta Electric System Operator. I would like to thank Akshaya Jha for outstanding research assistance.

Abstract

Several theoretical papers, most notably Borenstein, Bushnell and Stoft (2000), have demonstrated that transmission expansions can increase the amount of competition faced by wholesale electricity suppliers with the ability to exercise unilateral market. This perceived increase in competition faced by these strategic suppliers causes them to behave more aggressively and set market-clearing prices closer to competitive benchmark price levels. These lower wholesale market-clearing prices are the competitiveness benefit of this transmission policy to electricity consumers. This paper quantifies empirically for an actual wholesale electricity market the competitiveness benefits of a transmission expansion policy that causes strategic suppliers to perceive a very small frequency and duration of transmission constraints to limit the competition they face. Using hourly generation-unit level offer, output, market-clearing price and congestion data from the Alberta Wholesale Electricity Market from January 1, 2009 to December 31, 2011, this paper builds on the expected profit-maximizing offer model in Wolak (2003 and 2007) and best-reply offer pricing model in McRae and Wolak (2012) to compute two counterfactual no-perceived congestion (by the five largest strategic suppliers in Alberta) hourly market-clearing prices that are used to compute an upper and lower bound on the hourly competitiveness benefits of this transmission policy. Both competitiveness consumer benefits measures show economically substantial benefits from such a transmission policy. The lower bound approach which does not assume any actual transmission expansions, only a change in the perceived frequency of congestion, yields an average hourly consumer benefit of 3,067 Canadian Dollars (CAD). The upper bound which assumes that the perceived amount of congestion turns out to be the actual amount of congestion yields an average hourly consumer benefit of 79,590 CAD. Taken together, these empirical results argue in favor including competitiveness benefits in transmission planning processes in order to ensure that all transmission expansions with positive net benefits to electricity consumers are undertaken.

1. Introduction

The transition from a price-regulated, vertically-integrated regulated monopoly regime to the wholesale market regime in electricity supply industry has dramatically altered the role of the transmission network. Under the vertically-integrated monopoly regime, the price-regulated electric utility had a requirement to serve all demand in its service territory at the regulated price. This mandate provided a strong incentive for the utility to operate its existing generation units in a least-cost manner given the geographic location of daily electricity demand and make investments in additional transmission capacity when this was the least-cost approach to supply load growth in a given geographic area. In contrast, under the wholesale market regime the owner of the transmission network is financially independent of any generation unit owner and receives a regulated revenue stream that is largely independent of the level of congestion in the transmission network. An owner of multiple generation units selling into a wholesale market can therefore find it expected profit-maximizing to exploit the configuration of the transmission network to cause transmission congestion and shrink the size of the geographic market over which its units face competition in order to increase the revenues it receives from participating in the wholesale market.

For all of these reasons, the transmission network takes on a new role in the wholesale market regime as facilitator of competition. Specifically, the configuration of the transmission network determines the extent of competition that each supplier faces for a given geographic distribution of electricity demands. Transmission expansions can increase the number of hours of the year that a supplier faces sufficient competition to cause it to submit offer curves close to its marginal cost curve and thereby yield lower market-clearing prices than would be the case in the absence of the transmission expansion. Borenstein, Bushnell and Stoft (2000) use a two-node model of quantity-setting imperfect competition between two suppliers separated by finite-capacity transmission line serving price-responsive demands at both nodes to derive two theoretical results related to this question. First, limited transmission capacity between the two locations can give each firm an additional incentive to restrict its output in order to congest the transmission line into its local market in order to raise the price it receives for its output. Second,

relatively small investments in transmission capacity may yield significant increases in the competitiveness of realized market outcomes.[1]

The purpose of this paper is to quantify empirically the magnitude of the competitiveness benefits from transmission expansions for an actual wholesale electricity market. Several estimates are computed of the change in hourly short-term market prices and wholesale energy costs to consumers in the Alberta Wholesale Electricity Market (AWEM) that result from increasing the extent of competition that the five largest suppliers face because of a perceived reduction in the frequency and duration of transmission constraints. These counterfactual prices differ in terms of how the configuration of the transmission network is assumed to alter the extent of competition that these suppliers actually face. All of these counterfactual prices demonstrate economically significant competitiveness benefits to electricity consumers from a transmission policy that causes them to perceive a low frequency and duration of transmission constraints. These results imply that failing to account for this source of consumer benefits in the transmission expansion planning process for regions with formal wholesale electricity markets can leave transmission expansions with positive net benefits to electricity consumers on the drawing board.

The approach used to assess the competitiveness benefits of transmission expansions builds on the models of expected profit-maximizing offer behavior described in Wolak (2000, 2003, and 2007), where suppliers submit hourly offer curves into the short-term market to maximize their expected profits from selling energy given the distribution of residual demand curves they face. As shown in Wolak (2000), this residual demand curve distribution determines the extent of competition that a supplier faces, and therefore how close the supplier's offer curve is to its marginal cost curve. Transmission expansions typically flatted out the realized residual demand curves that a supplier faces because more offers from other locations in the transmission network are not prevented from competing with that supplier because of transmission constraints. These flatter residual demand curves cause an expected profit-maximizing supplier to submit an offer curve closer to its marginal cost curve. If all strategic suppliers face flatter

[1] Arrellano and Serra (2008) extend this result to the case of a cost-based short-term market similar to the ones in a number of Latin American countries. The amount of transmission capacity between the two regions impacts the mix of high fixed-cost and low variable cost base load capacity and low fixed-cost and high variable cost peaking capacity suppliers choose, with additional transmission capacity causing suppliers at both locations to choose a capacity mix closer to the socially efficient level.

residual demand curve realizations because of increased transmission capacity, then they will all submit expected profit-maximizing offer curves closer to their marginal cost which will yield market-clearing prices closer to competitive benchmark levels.

The major challenge associated with computing these counterfactual offer curves for each strategic supplier is quantifying how the curves will change in response to each supplier facing flatter residual demand curve distribution. The approach used here is based on framework implemented by McRae and Wolak (2012) to determine much a supplier's hourly offer prices (along its offer curve into hourly short-term market) changes in response to changes in the form of the hourly residual demand it faces. An econometric model relating the hourly offer price submitted by a supplier to the hourly inverse semi-elasticity of the residual demand curve (defined in McRae and Wolak (2012)) faced by that supplier is estimated for each of the five large suppliers in the AWEM using the hourly curves submitted by all market participants over the period January 1, 2009 to December 31, 2011. The hourly generation unit-level offer curves submitted by each of the five strategic market participants are used to compute each supplier's hourly offer price and the hourly market demand and aggregate offer curves of all other market participants are used to construct the hourly residual demand curve facing each strategic supplier.

This estimated relationship between the hourly offer price and hourly inverse semi-elasticity for each market participant is used to compute a counterfactual offer curve for each supplier that is the result of the perceived increased competition that the strategic supplier would face as result of increased transmission capacity. This is accomplished through the following three-step process. First, a no-congestion residual demand curve is computed for each hour for each supplier using the offer curves actually submitted by all suppliers. This residual demand curve assumes that the offer curves of all other suppliers, besides the firm under consideration, can compete against the offers of the firm under consideration. Second, the inverse semi-elasticity of this hourly no-congestion residual demand curve is computed and the coefficient estimates from the regression of the hourly offer price for that supplier on the actual hourly inverse semi-elasticity (that reflects actual transmission constraints) that the supplier faced is used to compute a counterfactual Canadian Dollar (CAD) per Megawatt-hour (MWh) reduction in the hourly offer price due to the smaller inverse semi-elasticity of the no-congestion residual demand curve. This CAD/MWh reduction is applied all the hourly offer prices for all steps on that supplier's offer curve. The final step of the process uses these counterfactual offer curves

for the five largest suppliers and the actual offer curves of the remaining suppliers to compute an aggregate counterfactual offer curve. The counterfactual hourly market price is computed by crossing the resulting aggregate offer curve with the actual demand for that hour. This three-step procedure is then repeated for all hours in the sample period.

The final step of the process is repeated in two ways in order to compute an upper and lower bound on the level of the counterfactual price that results from no perceived transmission constraints by the five large strategic suppliers. To compute a lower bound on the counterfactual no-congestion price (and upper bound on the economic benefits from transmission expansions), the counterfactual aggregate supply curve is computed using the adjusted offer curves for the strategic firms and actual offer curves for all other firms. The price at the intersection of this curve with the aggregate demand curve yield a lower bound on the counterfactual no-congestion price, because it assumes that there is sufficient transmission capacity that all of the offers on the aggregate offer curve below this counterfactual price can be accepted to supply energy.

To compute an upper bound on the counterfactual no-congestion price (and lower bound on the economic benefits from transmission expansions), the counterfactual aggregate supply curve is constructed using only quantity steps on the individual offer curves that were actually accepted. This implies that the counterfactual price is equal to the highest offer price with a positive quantity accepted from it in the actual hourly dispatch process. This second approach provides an extremely conservative estimate of the market price with no perceived transmission congestion because it assumes that exactly the same dispatch of generation units in the system and same amount transmission congestion as actually occurred. It is more likely to be the case that more the competitive behavior by strategic suppliers, even with same amount of transmission capacity, will allow some energy now offered at a lower price to sell energy and set a lower market-clearing price.

Both of these counterfactual prices indicate significant competitiveness benefits from transmission expansions that decrease the inverse semi-elasticity of the residual demand curve that the strategic supplier faces. These competitiveness benefits appear to correlated with the level of system demand for two reasons: (1) at high levels of system demand transmission constraints are more likely to limit the amount energy that compete against the strategic suppliers, and (2) at higher levels of the demand all suppliers typically face steeper residual demand curves even in the absence of transmission constraints because higher variable cost units

are needed to serve demand. In both cases, increasing the amount of transmission capacity increases the semi-elasticity of the residual demand curve (and decreases the inverse semi-elasticity) each suppliers faces, which our regression results imply will yield a lower offer price. There is also considerable variation in these competitiveness benefits across years in the sample, consistent with changes in the supply and demand balance over the three years of our sample.

The sample average hourly consumer benefit using the upper bound on the counterfactual no-perceived-congestion price is 3,067 CAD. However, this average hourly value varies considerably over the thirty-six months of the sample. During one month it exceeds 25,000 CAD. The sample average hourly competitiveness benefit using the lower bound on the counterfactual no-perceived-congestion price is 79,590 CAD. This magnitude also varies over months of the sample, taking on a value greater than 500,000 CAD for one month.

Translating these two consumer benefit measures from the perceived elimination of transmission constraints into percentages of the total cost of wholesale energy implies a lower bound on the consumer competitiveness benefits for the entire sample of 0.64 percent of total wholesale energy costs, with this percentage reaching has high as 2 percent of total wholesale energy costs in one month of the sample. For the entire sample, the upper bound on the competitiveness benefits is 16.8 percent of total wholesale energy costs. During a number of months, this percentage is substantially higher. For example, it is more than 45 percent of actual wholesale market revenues in one month. For most of the months this percentage is below 20 percent, but it never falls below 5 percent.

The remainder of this paper proceeds as follows. The next section describes the basic features of the AWEM and the process used to set market-clearing prices given the offers submitted to Alberta Electric System Operator (AESO). This section also presents summary statistics on the market structure and market outcomes in the AWEM. The third section describes the details of how the two counterfactual prices are computed. The fourth section presents the results of these computations. Section 5 discusses the implications of these results for the design of transmission planning processes in organized wholesale electricity markets.

2. The Alberta Wholesale Electricity Market

The AESO was formed in 2003 as a not-for-profit entity that is independent of all industry participants and owns no transmission or generation assets. It operates the AWEM,

which in 2011 had approximately 164 participants and processed close to $8 billion in electricity-related transactions. The AESO is governed by an independent board composed of members with backgrounds in finance, business, electricity, oil and gas, energy management, regulation, and technology development.

The AESO operates an hourly real-time energy market using a single-zone pricing model where one province-wide price of energy is set for each of hour of the day. Ancillary services are procured and dispatched by the AESO through an independent third-party market and over-the-counter transactions. The AESO dispatches these ancillary services to maintain adequate operating reserves throughout the day.

As shown in Table 1, thermal generation accounts for most of Alberta's energy production. Coal-fired generation accounts for slightly more than 46% of the installed capacity in Alberta. Natural gas-fired cogeneration is 27%, with natural gas-fired combined-cycle generation and natural gas-fired combustion turbine together accounting for slightly more than 11% of the installed capacity. The remaining capacity is wind, and biomass and other renewables. The dominant share of thermal capacity in the generation mix and significant differences in the variable cost across these generation technologies implies that there can be significant differences in the variable cost of the highest cost unit operating on the system throughout the day.

Table 1: Installed Capacity by Prime Mover

Prime Mover	Capacity in MW	Capacity Share (%)
Coal	6,232	46.29
Natural Gas Cogeneration	3,712	27.57
Hydroelectric	879	6.53
Natural Gas Combined Cycle	843	6.26
Wind	777	5.77
Natural Gas Combustion Turbine	753	5.59
Biomass and other renewables	266	1.98
Total Installed Capacity	13,462	100.00

The concentration of ownership of this generation capacity among suppliers to the Alberta market can influence the ability of suppliers to take unilateral actions to increase the profits they receive from selling energy into the AWEM. Table 2 lists the generation capacity controlled or owned by the five largest suppliers. These suppliers together control more than

three-quarters of the installed capacity in Alberta. The TransCanada controls almost 20% of the installed capacity, followed by TransAlta at 17.49%. The smallest of the five largest is Capital Power which controls more than 10% of the installed capacity in Alberta. This concentration of ownership of generation assets implies that high levels of fixed price forward contracts between generation unit owners and electricity retailers will be necessary to limit the incentive of these suppliers to exercise unilateral market power.

The benefits transmission expansions that cause of each of these suppliers to compete over the largest possible geographic market as many hours per year as possible are likely to be larger as a result of this concentration in generation capacity ownership. This logic implies that the competitiveness benefits of transmission expansions for this market are likely to substantial, even if suppliers have high-levels of hourly fixed-price forward contract obligations and therefore have limited incentives to exercise unilateral market power during most hours of the year. As shown in McRae and Wolak (2012), suppliers with hourly fixed price forward contract obligations close to the hourly output of their generation units have a significantly reduced incentive to take advantage of their ability to exercise unilateral market power. However, both unexpectedly high and unexpectedly low levels of output from a supplier's portfolio of generation units can create short periods when these suppliers have both the ability and incentives to exercise a significant amount of unilateral market power. A robust transmission network where transmission congestion is infrequent will limit the incentive to submit offer curves that reflect the exercise of substantial unilateral market power.

Table 2: Capacity Owned and Capacity Share of Five Largest Firms

Owner	**Capacity (MW)**	**Share of System (%)**
ATCO Power	1,349	16.52
Capital Power	1,507	11.19
ENMAX	1,897	14.09
TransAlta	2,354	17.49
TransCanada	2,580	19.17
Total of Five Largest Firms	9,687	78.46

Figure 1(a) plots the annual demand duration curves for the AWEM for 2009, 2010, and 2011. The highest recorded system peak demand is 10,609 MW. This was hit on January 16, 2012. System peaks in 2009, 2010, and 2011 were within a few hundred MWs of this level. The

horizontal axis on Figure 1(a) is the percentage of hours of the year from zero to 100 and the vertical axis is, from left to right, the hourly demand from the highest demand hour that occurred during the year to the lowest demand hour that occurred during the year. For a given percentage value on the horizontal axis, say 70 percent, MWh value on the vertical axis is the demand level that 70 percent of the hours of the year is above. Figure 1(a) shows that a significant amount of generation capacity is needed less than 5 percent of hours of the year. Figure 1(b) plots the curve for the 1 percent of the hours of the year with the highest hourly demands. For 2009, the difference between the annual peak demand and the demand at the highest 1^{th} percentile of the hourly demand distribution is almost 700 MWh. For 2010 and 2011, this difference is closer to 300 MWh.

Figure 2(a) plots the annual hourly price duration curves for 2009, 2010, and 2011. These curves are much flatter than the demand duration curves for all but the highest 15 percent of the hours of all three years. For the highest-priced 10 to 15 percent of the hours of the year, the curves become extremely steep, which is consistent with the earlier logic that the high levels of concentration of generation unit ownership can allow significant amounts of unilateral market power to be exercised during a small percentage of the hours of the year. Figure 2(b) plots the price duration curve for the highest 10 percent of hours of the year. For 2009 and 2010, this curve does not start to become steep until the highest 5 percent of hours of the year, whereas for 2011 this curve increases at close to a slope for the 10^{th} percentile to the highest priced hour of the year. For more than 10 percent of the hours of the year in 2011, prices are above 100 Canadian Dollars (CAD) per MWh. For 2009 and 2010, prices are above 100 CAD/MWh for approximately 5 percent of the hours of the year.

Transmission expansions that increase the competitiveness of the short-term market can also increase the incentive suppliers have to enter into fixed-price forward contract obligations. A supplier that faces greater competition more hours of the year as a result of increases in transmission capacity can create an additional incentive for that supplier to enter into a fixed-price forward contract that commits it to produce a higher level of output in the short-term market. This higher market-wide level of fixed-price forward contract coverage of final demand leads all suppliers to submit offer prices closer to their marginal cost of production, which yields market prices closer to competitive benchmark levels. These lower market prices are the primary source of benefits to electricity consumers from a higher capacity transmission network.

The analysis in this paper does not capture this forward contracting source of consumer benefits from transmission expansions. It only models the change in offer behavior brought about by each strategic supplier facing a more elastic residual demand curve because of the increased number of suppliers able to compete against it to supply energy because of the transmission expansion, not the potential change in that supplier's forward contracting decision and the forward contracting decisions of its competitors.

The consumer benefits of transmission expansions also depend of mechanism that translates the offer curves generation unit owners submit into the prices they are paid for the energy they produce. Generators in Alberta are able to set up to seven price and quantity pairs for each hour of the day for each generation unit in their portfolio. If (p_{ik},q_{ik}) i=1,2,3…,7 is the set of price level and quantity increment pairs for a generation unit k (k=1,2,..,K) owned by the supplier, that supplier's aggregate offer curve is constructed by ordering the offer price and quantity increment pairs from the lowest to highest offer price (regardless of generation unit) and then compute a step function with the height of each step equal to an offer price and the length of the step equal to the sum of the total amount of quantity increments across all generation units in that supplier's portfolio associated with that offer price. This yields the aggregate offer curve associated with that supplier.

Call the aggregate offer curve for supplier n during hour h, $S_h(p,\Theta_n)$, where Θ_n is the $14(K_n)$-dimensional vector of offer price and quantity increment pairs for the K_n generation units owned by supplier n. This curve gives the maximum amount of energy supplier n is willing to sell at price p during hour h. If there is no transmission congestion, then the market-clearing price is determined as the price where the aggregate supply curve intersects the aggregate demand during hour h, QD_h Mathematically, the market-clearing price, p^*, solves

$$S_h(p,\Theta_1) + S_h(p,\Theta_2) + , , , + S_h(p,\Theta_N) = QD_h, \quad (2.1)$$

where N is the total number of suppliers submitting offer curves during hour h.

` When there is transmission congestion that prevent the AESO from accepting a supplier's quantity increment, this quantity increment and its associated offer price is dropped from that supplier's offer curve. Define $SC_h(p,\Theta_n)$ as the transmission-constrained offer curve for supplier n during hour h. By definition of being transmission constrained, the following inequality holds for all price levels

$$SC_h(p,\Theta_n) \leq S_h(p,\Theta_n) \text{ for } p \quad (2.2)$$

and holds as a strict inequality for all prices greater than the lowest offer price at which a quantity increment cannot be accepted because of transmission constraints. Consequently, when there are transmission constraints, the market-clearing price, p^*, solves

$$SC_h(p,\Theta_1) + SC_h(p,\Theta_2) + , , , + SC_h(p,\Theta_N) = QD_h, \quad (2.3)$$

Figure 3 plots the aggregate offer curve not accounting for transmission constraints (called the Ideal Aggregate Offer Curve) and the offer curve with transmission constraints accounted for (called the Feasible Aggregate Offer Curve) for hour 12 of May 12, 2010. The vertical line in the graph is QD, the aggregate demand during that hour. The two curves satisfy inequality (2.2) for all prices from 0 to 1,000 CAD/MWh. Moreover, point of intersection of QD_h with the Ideal Aggregate Offer Curve yields a price that is much lower than the price at the intersection of the Feasible Aggregate Offer Curve, which determines the actual market-clearing price. The difference between the prices at the two points of intersection is almost 800 CAD/MWh. This price difference indicates the potential for significant consumer benefits from eliminating the transmission congestion that led to the need to use equation (2.3) to set the market-clearing price rather than equation (2.1).

If expected profit-maximizing suppliers believe that the transmission-constrained or Feasible Aggregate Offer Curve will be used to set prices rather than the unconstrained or Ideal Aggregate Offer Curve, these suppliers are likely to submit offer curves that make less capacity available at every output level relative to the case where they believe that the Ideal Aggregate Offer Curve will be used to set prices. The converse of this logic implies if each of the five large strategic suppliers believes that no quantity increment offers its competitors will be prevented from selling energy because of transmission constraints, then each strategic supplier will find it expected profit-maximizing to submit an offer curve closer to its marginal cost curve. This will yield lower market-clearing prices, whether or not some of its competitors' quantity increments are ultimately constrained from actually selling energy.

The next section describes how I estimate the change in each strategic supplier's offer curve in response to that supplier's belief that transmission constraints will not limit the competition that it faces for its output. The approach uses insights from the model of expected profit-maximizing offer behavior developed in Wolak (2000, 2003 and 2007). A methodology for computing both an upper bound and a lower bound on the "no-perceived-congestion" market-

clearing price that assumes no change in forward contracting behavior by the five large strategic suppliers is also derived.

3. Computing the "Perceived No-Congestion" Offer Curves and Counterfactual Market-Clearing Prices

This section summarizes the basic features of the model of expected profit-maximizing offer behavior introduced in Wolak (2000) and tested empirically in Wolak (2003 and 2007). This theoretical model and the empirical analysis in McRae and Wolak (2012) is the theoretical and empirical foundation for the procedure used to compute the "no-perceived-congestion" offer curve for each strategic supplier. These counterfactual offer curves and the actual offer curves of the remaining suppliers are used to compute the no-perceived-congestion counterfactual aggregate offer curves that are used to compute the counterfactual market prices associated with additional transmission capacity. Two counterfactual market-clearing prices are used to provide upper and lower bounds on the potential competitiveness benefits associated with a transmission network where congestion is expected to be infrequent.

My empirical modeling framework is based on the assumption that suppliers choose their offer curves to maximize the expected profits from selling energy given the distribution of aggregate demand and supply uncertainty and the offer curves chosen by their competitors. The offer curves of competitors and aggregate supply (primarily generation and transmission outages) and aggregate demand uncertainty creates a distribution of residual demand curve realizations that the supplier faces. As discussed in Wolak (2000), an expected profit-maximizing supplier picks the vector of parameters of its aggregate offer curve, Θ in the notation of the previous section, to maximize the expected value of the realized profits over the distribution of residual demand curves that it faces, subject to the constraints placed on the elements of Θ by the market rules. For example, in the AESO, all offer prices must be greater than or equal to zero and less than the offer cap, which is currently 1,000 CAD/MWh. The offer quantity increments must be greater than or equal to zero and their sum less than or equal to the capacity of the generation unit.

The price at the point of intersection of the supplier's offer curve with each residual demand realization determines the market-clearing price and amount of output that the supplier sells in the short-term market for that realization of residual demand uncertainty. This price and

quantity pair, along with the supplier's variable cost function, determines the supplier's realized variable profits for that residual demand realization. A described in detail in Wolak (2003 and 2007), an expected profit-maximizing supplier chooses the elements of Θ, the parameters of its offer curve, $S(p,\Theta)$, to maximize the expected value of these variable profit realizations with respect to the distribution of residual demand curve realizations.

It is important to emphasize that the assumption that suppliers maximize expected profits subject to the strategies of other market participants and the realizations of all supply and demand uncertainty is equivalent to that supplier exercising all available unilateral market power. A market participant is said to possess the ability to exercise market power if it can take unilateral actions to influence the market price and profit from the resulting price change. This means that the supplier faces a distribution of upward sloping residual demand curve realizations.

A shareholder-owned firm's management has a fiduciary responsibility to its shareholders to take all legal actions to maximize the expected profits it earns from participating in the wholesale market. Consequently, a firm is only serving its fiduciary responsibility to its shareholders when it exercises all available unilateral market power subject to obeying the wholesale market rules. A maintained assumption of our analysis is that both before and after a transmission upgrade, suppliers will choose their offer curves to maximize expected profits given the distribution of residual demand curves that they face. Consequently, if a transmission upgrade changes the distribution of residual demand curves that suppliers with the ability to exercise unilateral market power face, then the expected profit-maximizing offer curve each supplier submits should change. The remainder of this section describes how the change in offer behavior as a result of a reducing the incidence of transmission congestion is computed and how this change in offer behavior by the five strategic suppliers changes market-clearing prices.

3.1. Measuring the Ability to Exercise Unilateral Market Power in Bid-Based Markets

The residual demand curve that a supplier faces determines its ability to exercise unilateral market power. It is constructed from the offer curves submitted by all market participants besides the one under consideration. Let $S_n(p)$ denote the ideal offer curve of supplier n and $SC_n(p)$ the feasible offer curve of suppler n that accounts for transmission

constraints.[2] At each price, p, the function $S_n(p)$ gives the total quantity of energy that supplier n is willing to sell and the function $SC_n(p)$ gives the amount of energy supplier n is able to sell given the level and geographic location of demand, the offer curves submitted by its competitors and the configuration of the transmission network.

As shown in Figure 3, the offer curves from each supplier can be used to construct the Ideal Aggregate Offer Curve and the Feasible Aggregate Offer Curve. We can re-arrange equation (2.1) to derive the Ideal Residual Demand Curve for any supplier, which measures the ability of the supplier to exercise unilateral market in the absence of transmission constraints. To measure this ability of supplier j to exercise unilateral market power, equation (2.1) can be re-written as:

$$S_j(p) = QD - (S_1(p) + S_2(p) + ...+ S_{j-1}(p) + S_{j+1}(p) + ...+ S_N(p)) = QD - SO_j(p), \qquad (3.1)$$

where $SO_j(p)$ is the aggregate willingness-to-supply curve of all firms besides supplier j. Define $DR_j^I(p) = QD - SO_j(p)$ as the Ideal Residual Demand Curve facing supplier j. The ideal residual demand of supplier j at price p is defined as the market demand remaining to be served by supplier j after the ideal willingness-to-supply curves, $S_k(p)$ for all $k \neq j$ have been subtracted out.

The Feasible Residual Demand Curve facing supplier j can also be computed by re-arranging equation (2.3) in an analogous manner. This residual demand curve captures supplier j's ability to exercise unilateral market power given the actual configuration of the transmission network, location of demand and other generation units. In this case, equation (2.3) can be re-written as:

$$\begin{aligned} SC_j(p) &= QD - (SC_1(p) + SC_2(p) + ...+ SC_{j-1}(p) + SC_{j+1}(p) + ...+ SC_N(p)) \qquad (3.2) \\ &= QD - SCO_j(p), \end{aligned}$$

where $SCO_j(p)$ is the aggregate feasible willingness-to-supply curve of all firms besides supplier j. Define $DR_j^F(p) = QD - SCO_j(p)$ as the Feasible Residual Demand Curve facing supplier j. The feasible residual demand of supplier j at price p is defined as the market demand remaining to be served by supplier j after the feasible willingness-to-supply curves, $SC_k(p)$ for all $k \neq j$ have been subtracted out.

Equation (2.2) implies the following relationship between the Ideal and Feasible residual demand curves

[2]For simplicity, I have suppressed the dependence on Θ_k, the vector of price offers and quantity increments for supplier k.

$$DR_j^F(p) \geq DR_j^I(p) \text{ for all } p. \quad (3.3)$$

This relationship holds as a strict inequality for all prices greater than the lowest offer price associated with the first quantity offer from any firm besides supplier j that is prevented from being taken because of the configuration of the transmission network.

Figure 4(a) to 4(e) plot the Ideal and Feasible residual demand curves for the five largest suppliers in the Alberta market—ATCO Power, Capital Power, TransAlta, ENMAX, and TransCanada for hour 13 of May 16, 2010. The vertical line on each graph shows how much energy the supplier actually sold during that hour. For all five suppliers, the point of intersection between the Ideal Residual Demand Curve and the amount that the firm actually sold occurred at price that was substantially lower than price at which the Feasible Residual Demand curve intersected the amount the firm actually sold, which is also very close to the actual market-clearing price for that hour.

The expectation of facing a substantially steeper distribution of Feasible Residual Demand Curves would cause an expected profit-maximizing strategic supplier to submit a higher offer price for its output than it would if it faced the flatter distribution of Ideal Residual Demand Curves. Because I observe what offer curve each supplier actually submitted and what Feasible Residual Demand Curve it actually faced, using insights from the model of expected profit-maximizing offer behavior in Wolak (2000), I can follow the approach of McRae and Wolak (2012) to estimate the relationship between a supplier's hourly offer price and the form of the residual demand curve that it actually faced. This empirical relationship can then be used to estimate how the supplier's offer price would change in response to change in the form of the residual demand curve that it faced from the Feasible Residual Demand Curve to the Ideal Residual Demand Curve.

3.2. Measuring of the Ability to Exercise Unilateral Market Power from a Simplified Model of Expected Profit-Maximizing Offer Behavior

This section develops a simplified model of expected profit-maximizing offer behavior that motivates the linear regression model I estimate to predict how the hourly offer price of each of the five large strategic suppliers will change in response to facing the Ideal Residual Demand Curve for that hour rather than the Feasible Residual Demand Curve for that hour. This linear regression model has been employed by McRae and Wolak (2012) to predict how strategic suppliers in the New Zealand wholesale electricity market will change their half-hourly offer

prices in response to changes in the form of the half-hourly residual demand curve they face. McRae and Wolak (2012) found that even after controlling for differences input fuel costs across days of their sample, when each of the four large New Zealand suppliers faced less competition, as measured by the half-hourly value of the inverse semi-elasticity of their residual demand curve, each of the firms was predicted to submit a significantly higher half-hourly offer price.

Although a supplier does not know with certainty the market demand and the willingness-to-supply offers of other suppliers when it submits its offers, the supplier does have a very good idea of the set possible realizations of the residual demand curves it might face. The characteristics of each generation unit owned by the supplier's competitors and the market rules can significantly constrain the set of offers curves a supplier can submit. The pattern of hourly electricity demands throughout the day is very similar across weekdays within the same season of the year. In addition, all market participants understand the impact of weather conditions on the demand for electricity and the likely availability of intermittent resources like hydroelectric energy and wind energy. Finally, all suppliers monitor the daily prices of the fossil fuel inputs and the availability of these inputs.

All of these factors imply that a large supplier has a very good idea about the set of possible residual demand curve realizations that it might face. For each possible residual demand curve realization the supplier can find the ex post profit-maximizing market price and output quantity pair given its marginal cost curve following the process described above. This is the market price and output quantity pair that the supplier would like to achieve for that residual demand curve realization.

Figure 5(a) illustrates the construction of an expected profit-maximizing willingness to supply curve using this process for the case of two possible continuously differentiable residual demand curve realizations. For each residual demand curve realization, intersect the marginal cost curve with the marginal revenue curve associated with that residual demand curve realization. For example, for Residual Demand Curve 1 the marginal revenue curve for this residual demand curve (not shown on the figure) intersects the marginal cost curve at the quantity Q_1. The output price associated with this output level on Residual Demand Curve 1 is P_1. Repeating this process for Residual Demand Curve 2 yields the profit-maximizing price and quantity pair (P_2,Q_2). Note that because both residual demand curves are very steeply sloped, there is a substantial difference between the market price and the marginal cost at each output

level. If these two residual demand realizations were the only ones that the supplier faced, its expected profit-maximizing offer curve would pass through both of these points because regardless of the residual demand realization this offer curve would cross at an ex post expected profit-maximizing level of output. The straight line connecting the points (P_1,Q_1) and (P_2,Q_2) in the figure is one possible expected profit-maximizing offer curve.

To illustrate the impact of more elastic residual demand curves on the offer curves submitted by an expected profit-maximizing supplier, Figure 5(b) repeats the construction of an expected profit-maximizing offer curve for the case of two more elastic residual demand curve realizations. The line connecting the points (P_1,Q_1) and (P_2,Q_2), which is an expected profit-maximizing offer curve for these two residual demand realizations, is much closer to the supplier's marginal cost curve. Specifically, for each residual demand realization, the price associated with the profit-maximizing level of output for that residual demand curve realization is closer to the marginal cost of producing that level of output than it was in Figure 5(a). This outcome occurs because each residual demand realization is much more elastic than the residual demand realizations in Figure 5(a).

Figure 5(c) considers the case in which the two residual demand curve realizations are infinitely elastic, meaning that for each realization the supplier faces enough competition that the entire market can be satisfied at a fixed price. By the logic described above, the supplier will find it unilaterally profit-maximizing to produce at the intersection of each residual demand curve realization with its marginal cost curve, because the marginal revenue curve for each residual demand realization is equal to the residual demand curve. In this case, the supplier's expected profit-maximizing offer curve, the line connecting the profit-maximizing output levels for each residual demand curve realization, is equal to the supplier's marginal cost curve. This result illustrates a very important point that if a supplier faces sufficient competition for all possible residual demand curve realizations then it will find it unilaterally expected profit-maximizing to submit an offer curve equal to its marginal cost curve.

The examples in Figures 5(a) to 5(c) utilize continuously differentiable residual demand curves. However, the same process can be followed to compute an expected profit-maximizing offer curve for the case of step-function residual demand curves. Figure 5(d) shows how this would be done for the more realistic case of step function residual demand curves with two possible residual demand realizations. For each residual demand curve realization, the supplier

would compute the ex post profit-maximizing level of output and market price for the marginal cost curve given in Figure 5(d). For DR_1 this is the point (P_1,Q_1) and for DR_2 this is the point (P_2,Q_2). If these two residual demand curve realizations were the only possible residual demand curve realizations that the supplier could face, then a step function offer curve that passes through these two points would be an expected profit-maximizing offer curve.

Computing the expected profit-maximizing offer curve for a supplier is generally more complex than passing an offer curve through the set of ex post expected profit-maximizing price and output quantity pairs every possible residual demand curve realization. That is because the market rules can prevent a supplier from achieving the ex post profit-maximizing market price and output quantity pair for all possible residual demand realizations. Specifically, unless all of these ex post profit-maximizing price and quantity pairs lie along a willingness-to-supply curve for the supplier that the market rules allow it to submit, it is not possible for the supplier to submit a willingness to supply curve that always crosses the realized residual demand curve at an ex post profit-maximizing price and quantity pair for that residual demand curve realization.

Figure 5(e) provides an example of this phenomenon. This figure shows the ex post profit-maximizing price and quantity pairs for three residual demand curves. Note that the profit maximizing point for DR_3 lies below and to the right of the profit maximizing point for DR_1. This makes it impossible for the supplier to submit a non-decreasing step function offer curve that passes through the three ex post profit-maximizing price and output quantity pairs. In this case, the supplier must know the probability of each residual demand curve realization in order to choose the parameters of its expected profit-maximizing willingness to supply curve.

Figure 5(e) demonstrates that the expected profit-maximizing residual demand curve does not pass through any of these three ex post profit-maximizing price/quantity pairs. Instead, as discussed in Wolak (2003 and 2007), the form of the expected profit-maximizing willingness-to-supply curve depends on both the form of each residual demand curve realization and the probability of that residual demand curve realization. This curve, shown in Figure 5(e), yields market-clearing price and quantity-sold pairs for the firm for each of the three residual demand curve realizations that maximize the expected profits the firm earns subject to this offer curve being in the set of offer curves the market rules allow a supplier to submit. As shown in Wolak (2003) and Wolak (2007), the supplier chooses the price level and quantity increments that

determine its offer curve to maximize its expected profits over the distribution of possible residual demand curve realizations that it faces.

The basic intuition from the continuously differentiable residual demand curve analysis also holds for the general case of step function residual demand curve. When a supplier faces a flatter distribution of residual demand realizations, it will find it expected profit-maximizing to submit a willingness-to-supply curve with offer prices closer to its marginal cost of production. Following McRae and Wolak (2012), I use the simplified model of expected profit-maximizing offer behavior to derive a summary measure of the hourly unilateral ability of a supplier to exercise market power from the realized residual demand curve that the supplier faced during that hour. This measure, called the Inverse Semi-Elasticity of the realized residual demand curve at the actual market-clearing price provides an ex post measure of the ability of a supplier to exercise unilateral market power. Specifically, this inverse semi-elasticity quantifies the $/MWh increase in the market-clearing price that would have occurred if the supplier had reduced the amount of output it sold in the market by one percent. This interpretation of the inverse semi-elasticity of the residual demand curve does not rely on the assumption that the realized output level and market-clearing price maximize the supplier's ex post profits as is the case for the continuously differentiable residual demand curve realizations in Figures 5(a) to 5(c).

As shown in McRae and Wolak (2012), the simplified model of expected profit-maximizing offer behavior described in Figures 5(a) to 5(c), implies a linear relationship between the offer price along the supplier's offer curve, its marginal cost of production and the inverse semi-elasticity of the realized residual demand curve. The first-order conditions for ex post profit-maximization for these two residual demand realizations in Figure 5(a) imply:

$$P_i = C_i - [DR_i(P_i)/DR_i'(P_i)], \quad i=1,2. \tag{3.4}$$

Equation (3.4) implies that the offer price for the supplier at its output level for residual demand curve realization 1 or 2 (P_i for i=1,2) is equal to the marginal cost of the highest cost unit owned by that supplier operating for that residual demand curve realization (C_i for i=1,2) plus the value of the residual demand curve at that offer price divided by the absolute value of the slope of the residual demand curve at that offer price for the residual demand curve realization ($[DR_i(P_i)/DR_i'(P_i)]$ for i=1,2).

Define η_i (i=1,2) as the inverse semi-elasticity of the residual demand curve i, as:

$$\eta_i = - (1/100)[DR_i(P_i)/DR_i'(P_i)]. \tag{3.5}$$

at offer price P_i for i=1,2. This magnitude gives the $/MWh increase in the market-clearing price associated with a one percent reduction in the amount of output sold by the supplier. In terms of this notation, equation (3.4) becomes

$$P_i = C_i + 100\eta_i, \; i=1,2. \tag{3.6}$$

Thus, the simplified model of expected profit-maximizing offer behavior implies that higher hourly offer prices for the supplier should be associated with higher values of the hourly inverse semi-elasticity.

As discussed above, because offer curves in the AWEM are step functions, defining a value of η_i, the inverse semi-elasticity, for a step function residual demand curve requires choosing a method for computing a finite difference approximation to the slope of the residual demand curve at a specific value of the offer price. This logic also implies that because actual residual demand curves are step functions, equation (3.6) will not hold with equality for the computed values of the inverse semi-elasticity. However, the general model of expected profit-maximizing offer behavior with step function offer curves and residual demand curves described above implies that when a supplier has a greater ability to exercise unilateral market power as measured by the size of η_i, that supplier's offer price is likely to be higher. Wolak and McRae (2011) presented empirical evidence consistent with this hypothesis for the four largest suppliers in the New Zealand wholesale electricity market.

The method for calculating the finite difference slope of the step-function residual demand curve at the firm's actual hourly output level requires choosing the output change used to compute the finite-difference approximation to the slope. These output changes should be large enough to ensure that price steps on the residual demand curve are crossed so that a non-zero slope is obtained, but not too large that the implied output change is judged implausible for the supplier to implement. McRae and Wolak (2012) experimented with a number of approaches to computing this finite difference approximation to the slope and found their empirical results were largely invariant to the approach used. I follow their preferred approach to computing the finite difference slope of the residual demand curve that enters into the computation of the hourly inverse semi-elasticity of the residual demand curve for each strategic supplier.

3.3. The Counterfactual No-Perceived-Transmission-Constraints Offer Curve

This section describes how I compute the counterfactual offer curve for each strategic supplier under the assumption of no perceived transmission constraints, which means that the

strategic suppliers expect to face the Ideal Residual Demand Curve rather than the Feasible Residual Demand Curve. I first compute the hourly inverse semi-elasticity of the Feasible Residual Demand curve facing each strategic supplier for the entire sample period. Then for each strategic supplier, I compute a linear regression analogue of equation (3.6) where the supplier's hourly offer price at its actual output level for that hour is regressed on day-of-sample and hour-of-day fixed effects (that control for across-day changes in input prices and within-day variation in operating costs) and the hourly inverse semi-elasticity of the Feasible Residual Demand Curve faced by that supplier.

The coefficient estimate on the hourly inverse semi-elasticity is used to compute the predicted change in the supplier's offer price as a result of facing the Ideal Residual Demand Curve instead of the Feasible Residual Demand Curve. This offer price change is applied to all offer prices along that firm's willingness-to-supply curve. The process is repeated for all hours of the sample period to compute a counterfactual no-perceived-congestion offer curve for each hour of the sample period. This process is then repeated for all strategic suppliers.

The second column of Tables 3(a) to 3(e) lists the daily averages of the inverse semi-elasticities of the Feasible Residual Demand Curve for hour h for supplier (n=ATCO, Capital Power, ENMAX, TransAlta, and TransCanada), η_{nh}^{F}, for each hour of the day over the sample period January 1, 2009 to December 31, 2011. The third column in each table lists the daily averages of the inverse semi-elasticities for Ideal Residual Demand Curve for the hour h, η_{nk}^{I}, for the same five suppliers for each hour of the day over the sample period January 1, 2009 to December 31, 2011. Note that consistent with the inequality in (3.3) the sample mean of η_{nh}^{F} is greater than the sample mean of η_{nk}^{I} for all hours of the day for all five strategic suppliers. The differences are much larger during the peak demand hours of the day when transmission constraints are likely render more quantity offers unable to be accepted to supply energy. This result is consistent with more of the competitiveness benefits of transmission investments being realized during the high demand hours of the day, week, and year.

In order to describe the linear regression analogue to equation (3.6) that I estimate to predict changes in each strategic supplier's offer price as result of facing the Ideal Residual Demand Curve rather than the Feasible Residual Demand Curve, a definition of a supplier's hourly offer price is required. Figure 6 presents the actual hourly offer curve for a hypothetical Firm A. The dispatched quantity of energy for Firm A during that hour is 1,508 MW. The offer

price along Firm A's willingness-to-supply curve for that hour is found by extending a vertical line up from the horizontal axis at 1,508 MW until it intersects Firm A's willingness-to-supply curve. In this case, the offer price for the dispatched quantity for Firm A is equal to $145/MWh, which is the offer step directly above the quantity level 1,508 MW. In general, the offer price for output level Q^* for supplier k during hour h is computed as the solution to the following equation in P: $Q^* = S_{hn}(P)$, where $S_{hn}(P)$ is supplier n's willingness-to-supply curve during hour h.

Equation (3.6) from the simplified model of expected profit-maximizing offer behavior by a supplier facing a distribution of downward sloping continuously differentiable residual demand curves implies that,

$$P_{hn} = C_{hn} + \beta\eta_{hn}^{F}, \tag{3.7}$$

where P_{hn} is the offer price of supplier n during hour h, C_{hn} is the marginal cost of the most expensive generation unit owned from supplier n that is operating during hour h, and η_{hn}^{F} is the inverse semi-elasticity of the Feasible Residual Demand Curve of supplier n during hour h, and β is an unknown parameter to be estimated. Equation (3.7) implies that after controlling for the opportunity cost of the highest cost generation unit operating during that hour, C_{hn}, a supplier's offer price at the quantity of energy that it sells in the short-term market should be an increasing function of the value of the inverse semi-elasticity.

Let $P_{jhdm}(\text{offer})$ equal the offer price at the actual level of output sold by supplier j during hour h of day d during month of sample m, η_{jhdm}^{F}, the inverse semi-elasticity of supplier j's Feasible Residual Demand Curve during hour h of day d during month of sample m. I control for differences across hours during our sample period in the variable cost of the highest cost generation unit owned by that supplier operating during hour h by allowing for day-of-sample fixed effects and hour-of-day fixed effects for each supplier. The following regression is estimated for each supplier j:

$$P_{jhdm}(\text{offer}) = \alpha_{dmj} + \tau_{hj} + \beta_j\eta_{jhdm} + \varepsilon_{jhdm}, \tag{3.8}$$

where the α_{dmj} and γ_{dmj} are day-of-month d and month of sample m fixed effects and the τ_{hj} and are hour-of-the-day fixed effects for supplier j. The ε_{jhdm} are mean zero and constant variance regression errors.

\

Table 3(a): Daily Means of Hourly Feasible and Ideal Inverse Semi-Elasticites for ATCO

Hour	Feasible Inverse Semi-elasticity	Ideal Inverse Semi-elasticity
0	6.3913	4.5223
1	4.5759	3.0814
2	4.1091	2.5671
3	3.1988	1.7622
4	3.7905	2.1794
5	4.0035	2.3963
6	9.2222	6.283
7	29.2169	24.9878
8	21.7098	11.7629
9	41.5394	33.8444
10	41.1473	29.7382
11	50.5034	28.8659
12	29.4344	19.192
13	42.7524	21.8586
14	29.8361	19.1386
15	51.7264	30.8415
16	56.1854	33.2363
17	79.4979	54.4671
18	52.3705	29.9049
19	35.7296	17.0578
20	37.1703	29.5469
21	28.7581	16.9043
22	11.0723	6.3052
23	9.0169	4.5579

Table 3(b): Daily Means of Hourly Feasible and Ideal Inverse Semi-Elasticities for Capital Power

Hour	Feasible Inverse Semi-elasticity	Ideal Inverse Semi-elasticity
0	6.9569	4.7097
1	4.8931	3.0866
2	4.5806	2.6925
3	4.427	1.9455
4	5.2092	2.3134
5	4.7518	2.494
6	9.9295	6.5786
7	29.1587	24.9033
8	22.5017	12.8093
9	41.8103	33.5288
10	30.1607	22.3046
11	51.2066	28.7834
12	27.7195	17.8637
13	40.7075	28.6439
14	28.6991	19.2485
15	47.7725	36.331
16	59.0699	30.445
17	65.9477	49.8988
18	69.1007	32.5257
19	31.3424	14.4765
20	57.4056	30.3224
21	26.4011	13.5428
22	10.4818	6.378
23	8.7793	4.76

Table 3(c): Daily Means of Hourly Feasible and Ideal Inverse Semi-Elasticites for ENMAX

Hour	Feasible Inverse Semi-elasticity	Ideal Inverse Semi-elasticity
0	9.4245	5.4906
1	7.0398	3.7746
2	6.2446	3.1617
3	5.1674	2.4551
4	7.2173	3.0502
5	8.8495	4.5622
6	10.7127	6.3759
7	23.544	19.1729
8	25.4544	14.6554
9	39.3389	32.0813
10	33.5831	23.8562
11	42.7017	22.2442
12	30.062	17.0049
13	38.7434	20.3458
14	37.0706	22.4362
15	39.0619	25.1733
16	51.8622	31.265
17	54.6498	45.0149
18	58.7102	24.103
19	34.7564	15.9928
20	39.4	28.7643
21	28.7876	17.1298
22	12.1551	7.735
23	10.2982	5.808

Table 3(d): Daily Means of Hourly Feasible and Ideal Inverse Semi-Elasticities for TransAlta

Hour	Feasible Inverse Semi-elasticity	Ideal Inverse Semi-elasticity
0	6.3966	4.5438
1	4.3691	2.8889
2	4.0839	2.59
3	3.3936	1.9453
4	3.6537	2.1542
5	4.0189	2.3412
6	11.5959	7.2019
7	29.6453	25.3327
8	24.3716	13.4542
9	41.5292	33.4996
10	28.4284	21.3949
11	53.8608	34.1311
12	24.1888	15.9755
13	37.0122	21.3211
14	26.4141	16.9625
15	39.6917	28.9546
16	49.7388	31.4801
17	60.6202	48.4901
18	57.5477	26.639
19	27.3654	15.7026
20	34.4431	27.0927
21	23.3192	12.2585
22	9.9137	6.0877
23	8.0715	4.7127

Table 3(e): Daily Means of Feasible and Ideal Hourly Inverse Semi-Elasticities for TransAlta

Hour	Feasible Inverse Semi-elasticity	Ideal Inverse Semi-elasticity
0	13.5245	8.1049
1	8.6496	4.1487
2	13.5271	7.8086
3	9.0608	4.2187
4	10.877	4.2615
5	12.7985	7.0101
6	18.0116	11.7573
7	39.0842	31.6787
8	27.4971	16.1518
9	60.5141	35.5801
10	57.4434	28.3626
11	53.2884	32.6641
12	38.4105	21.6951
13	44.8432	24.5349
14	63.6096	22.1926
15	63.5564	43.7651
16	72.6378	40.3488
17	87.5616	61.5054
18	76.4675	37.6527
19	41.8704	19.8875
20	42.128	32.6894
21	34.3374	15.5131
22	14.2099	7.9596
23	12.8844	6.3362

These fixed effects control for variation in costs and operating conditions and across days of the sample and within days. Input fossil fuel prices and hydroelectric water levels change at most on a daily basis. Because there is a different fixed effect for each day and month combination during our sample period, these fixed effects completely account for the impact of daily changes in fossil fuel prices and water levels during our sample period on the variable cost of the highest cost generation unit owned by supplier j that is operating during each hour during the day. For these reasons, these day-of-sample fixed-effects completely account for any day-to-day change in the prices of input fossil fuels such as natural gas and coal paid by supplier j. The hour-of-day fixed-effects account for differences across hours of the day in the variable cost of the highest cost generation unit in that supplier's portfolio operating. This strategy for controlling for variable cost changes across hours of the sample implies that more than 1,100 parameters determine the hourly variable cost values for each supplier over the sample period Multiplying this figure by five implies more than 5,500 parameters determine the hourly variable cost of the highest cost generation unit operating during a hour of sample across the five strategic suppliers. For all of these reasons, the day-of-sample and hour-of-day fixed effects for all five strategic suppliers should be more than sufficient to account for changes in the variable cost of the highest cost unit operating during hour h of day d of month of sample m.

Table 4 presents the estimated values of β_j and the estimated standard errors for each of the five largest suppliers from estimating equation (3.8) for each supplier over our sample period of January 1, 2009 to December 31, 2011. The values of β_j are positive, precisely estimated and economically meaningful for all regressions.

Table 4: Coefficient Estimates of β_j in Regression (3.8) for Supplier j

	Coefficient Estimate	Standard Error
$\beta_{TransAlta}$	0.0552976	0.002048
$\beta_{TransCanada}$	0.0546614	0.00137
β_{ENMAX}	0.0423829	0.001491
$\beta_{Capital\ Power}$	0.0574105	0.001564
$\beta_{ATCO\ Power}$	0.097143	0.002224

Note: Each line of the table corresponds to a different regression with 1,095 day-of-sample and 24 hour-of-day fixed effects included in each regression.

Each of these regression coefficient estimates implies that holding all other factors constant, if the inverses semi-elasticity of the residual demand curve faced by one of the five large suppliers falls, then the offer price for that firm is predicted to fall by the change in the

semi-elasticity times the estimated value of β_j for that supplier. Tables 5(a) to 5(e) contain the hourly sample standard deviations of the hourly Feasible and Ideal inverse semi-elasticities. The standard deviations for the Feasible inverse semi-elasticities are in the range of 300 to 600 CAD/MWh during a number of hours of the day for each of the suppliers. This implies that a one standard deviation increase in the hourly inverse elasticity for one of these hours of the day predicts an increase in supplier's hourly offer price of 15 to 40 CAD/MWh for the regression coefficient estimates in Table 4.

This result indicates that the potential for economically significant competitiveness benefits from transmission expansions that reduce both the mean and standard deviation of the hourly inverse semi-elasticities. The standard deviations of the Ideal inverse semi-elasticities are uniformly smaller than the corresponding values for the Feasible inverse semi-elasticities. This result demonstrates an additional source of competitiveness benefits from transmission expansions that reduce the frequency and magnitude of congestion. These expansions reduce the incidence of extremely large inverse semi-elasticities which the results in Table 4 imply will lead to substantially larger offer prices and substantially larger market-clearing prices.

The final step in the process of computing the counterfactual no-perceived-congestion offer curve adjusts each offer price submitted by supplier j during hour h by the difference between the Feasible semi-elasticity and the Ideal semi-elasticity times the estimated value of β_j. Mathematically, if P_{jhk} is the offer price for bid quantity increment k for supplier j during hour h, then the no perceived congestion offer price for this bid quantity increment is:

$$P_{jhk}^{NC} = P_{jhk} - \beta_j(\eta_{hn}^{F} - \eta_{hn}^{I}) \; . \qquad (3.9)$$

Repeating this process for all bid quantity increments yields a new vector of offer price and quantity increment pairs, Θ^{NC}. This vector is composed of the modified offer prices, P_{jhk}^{NC}, from (3.9) and original offer quantity increments. Let $S_h(\Theta_n^{NC})$ denote the modified no perceived congestion offer curve for the supplier n during hour h.

Figure 7(a) to 7(c) illustrate the process used to compute $S_h(\Theta_n^{NC})$, from $S_h(\Theta_n)$, original offer curve for supplier n during hour h for hypothetical Firms A and B. The upper step function in Figures 7(a) and 7(b) are the original willingness-to-supply curves for Firms A and B. The lower step functions in the figures are the shifted down no-perceived congestion willingness-to-supply curves of Firms A and B. The upper step function in Figure 7(c) is the original aggregate willingness-to-supply curve of Firms A and B and the lower step function is

the shifted no-perceived-congestion aggregate willingness-to-supply curve for the two firms. Figure 7(c) demonstrates that for the same level of aggregate demand, the shifted no-perceived-congestion aggregate willingness-to-supply curve will set a lower market-clearing price than the original aggregate willingness-to-supply curve. This market price reduction is the source of the competitiveness benefits to electricity consumers from transmission investments.

4. The Competitiveness Benefits of Congestion-Reducing Transmission Investments

This section describes the calculation of the two counterfactual no-perceived-congestion market-clearing prices. The results of computing these two prices for all hours from January 1, 2009 to December 31, 2011 are described and then several calculations are presented to demonstrate the magnitude of consumer benefits from transmission expansions that reduce the frequency and magnitude of transmission congestion.

The first counterfactual price takes an extremely conservative approach to computing the competitiveness benefits of transmission expansions. It assumes no change in what offer quantities can be accepted because of transmission constraints. The only difference is that the Feasible Offer Curve for the five large strategic suppliers uses the adjusted offer prices from equation (3.9). In terms of the notation of Section 3, the offer curves for the strategic suppliers are defined as $SC_h(\Theta_n^{NC})$, the Feasible Offer Curve defined in Section 2 evaluated at Θ_n^{NC}, instead of Θ_n. This counterfactual price provides a very slack upper bound on market-clearing price that would result if all strategic suppliers faced the Ideal Residual Demand curve instead of the Feasible Residual Demand curve.

Table 5(a): Daily Standard Deviations of Hourly Feasible and Ideal Inverse Semi-Elasticites for ATCO

Hour	Feasible Inverse Semi-Elasticity	Ideal Inverse Semi-Elasticity
0	41.561	40.743
1	15.322	14.893
2	13.31	12.571
3	9.364	7.605
4	10.295	7.981
5	10.924	9.633
6	40.067	35.372
7	292.143	290.078
8	132.398	60.35
9	343.616	311.174
10	253.786	235.329
11	292.157	164.427
12	98.242	74.311
13	274.736	102.382
14	134.156	97.913
15	359.722	215.827
16	271.568	206.438
17	456.373	308.841
18	280.885	227.526
19	254.419	71.31
20	235.152	226.993
21	162.497	110.504
22	48.755	32.261
23	44.892	18.773

Table 5(b): Daily Standard Deviations of Feasible and Ideal Hourly Inverse Semi-Elasticities for Capital Power

Hour	Feasible Inverse Semi-elasticity	Ideal Inverse Semi-Elasticity
0	42.075	41.232
1	15.016	14.058
2	15.484	12.636
3	29.114	7.747
4	29.606	7.938
5	13.884	9.509
6	40.304	35.644
7	291.856	289.94
8	133.435	64.155
9	342.833	310.167
10	119.003	106.736
11	283.955	170.37
12	93.317	72.353
13	254.823	235.151
14	114.195	98.2
15	316.323	302.876
16	337.785	202.272
17	322.559	286.61
18	611.637	289.342
19	158.011	59.6
20	469.176	234.132
21	136.791	60.309
22	34.966	32.301
23	30.378	18.867

Table 5(c): Daily Standard Deviations of Hourly Feasible and Ideal Inverse Semi-Elasticites for ENMAX

Hour	Feasible Inverse Semi-elasticity	Ideal Inverse Semi-elasticity
0	54.056	42.118
1	25.593	16.602
2	23.678	12.669
3	15.524	8.921
4	31.927	11.829
5	43.152	33.979
6	40.43	31.318
7	194.786	191.771
8	144.113	77.799
9	320.685	286.365
10	195.535	125.898
11	236.523	106.898
12	104.398	68.418
13	235.702	80.995
14	177.397	151.047
15	202.042	164.03
16	290.463	204.292
17	355.793	353.725
18	548.399	201.402
19	221.705	57.52
20	232.545	221.435
21	151.768	90.505
22	42.209	40.508
23	32.312	22.947

Table 5(d): Daily Standard Deviations of Feasible and Ideal Hourly Inverse Semi-Elasticities for TransAlta

Hour	Feasible Inverse Semi-elasticity	Ideal Inverse Semi-elasticity
0	41.834	41.227
1	14.433	14.01
2	13.116	12.704
3	11.299	9.914
4	9.708	7.875
5	10.881	9.48
6	96.537	48.327
7	296.072	292.202
8	155.372	75.309
9	343.144	310.183
10	116.518	104.384
11	328.037	239.922
12	84.439	68.17
13	251.092	100.32
14	111.537	89.703
15	232.572	212.737
16	288.604	209.807
17	300.503	284.992
18	555.296	222.229
19	127.804	64.702
20	227.523	219.968
21	131.424	59.054
22	35.701	32.304
23	29.438	20.325

Table 5(e): Daily Standard Deviations of Feasible and Ideal Hourly Inverse Semi-Elasticities for TransAlta

Hour	Feasible Inverse Semi-elasticity	Ideal Inverse Semi-elasticity
0	65.299	59.51
1	26.253	12.322
2	125.547	116.741
3	43.99	17.537
4	54.178	19.225
5	61.63	51.984
6	79.64	71.81
7	320.28	313.171
8	140.397	73.597
9	497.988	311.275
10	485.63	133.104
11	291.225	204.085
12	167.336	84.217
13	250.864	88.981
14	774.835	96.416
15	362.656	320.241
16	429.877	321.408
17	400.863	335.92
18	630.742	345.018
19	253.976	91.884
20	312.371	305.789
21	174.459	77.193
22	37.217	31.736
23	34.718	17.451

To compute this counterfactual price for hour h, $SC_h(\Theta_n^{NC})$ is used for each of the five large strategic suppliers and the original feasible offer curve is used for all other suppliers. For simplicity assume that n=1,2,..,5 corresponds the five strategic firms and the remaining non-strategic firms are indexed n=6,7,..,N. The first counterfactual no-perceived transmission congestion market-clearing price for hour h is computed by solving for the smallest price such that:

$$SC_h(p,\Theta_1^{NC}) + SC_h(p,\Theta_2^{NC}) + , , , + SC_h(p,\Theta_5^{NC}) + SC_h(p,\Theta_6), + \ldots, + SC_h(p,\Theta_N) = QD_h, \quad (4.1)$$

Because the highest offer price accepted during h could be from a non-strategic firm, even though all of the adjusted offer prices of the strategic suppliers in Θ_n^{NC} (n=1,2,…,5) are less than the original offer prices in Θ_n (n=1,2,…,5), this market-clearing price, PC_h^F, is less than or equal to the actual market-clearing price, P_h. This weak inequality holds as a strict inequality unless the offer price of a non-strategic firm set the original market-clearing price.

To compare this Feasible Offer Curve counterfactual price-setting process to the actual hourly price-setting process used by the AESO, I also compute an estimate of the actual market-clearing price using the original Feasible Offer Curves of all suppliers. Let PP_h^F denote the smallest price that solves:

$$SC_h(p,\Theta_1) + SC_h(p,\Theta_2) + , , , + SC_h(p,\Theta_5) + SC_h(p,\Theta_6), + \ldots, SC_h(p,\Theta_N) = QD_h, \quad (4.2)$$

Note that original offer price and feasible offer quantities are used in the Feasible Offer Curves of all suppliers to compute the Predicted Feasible Actual market-clearing price, PP_h^F.

Figure 8 plots the daily average value of the actual market-clearing price and the daily-average of the Predicted Feasible Actual market-clearing price. In spite of the fact that daily average of actual prices is extremely volatile, sometimes exceeding 600 CAD/MWh, the daily average of the Predicted Feasible Actual market-clearing price is virtually identical for days of the sample period from January 1, 2009 to December 31, 2011.

The second counterfactual no-perceived congestion market-clearing price yields a lower bound on the no-perceived-congestion counterfactual price. It assumes that all suppliers face no transmission constraints so that the counterfactual market-clearing price is computed from the Ideal Offer Curves of the five strategic suppliers using the offer prices adjusted as described in equation (3.9) and the Ideal Offer Curves of the non-strategic suppliers. Mathematically, the counterfactual no-perceived congestion price, PC_h^I, is the smallest price that solves:

$$S_h(p,\Theta_1^{NC}) + S_h(p,\Theta_2^{NC}) + , , , + S_h(p,\Theta_5^{NC}) + S_h(p,\Theta_6), + \ldots, + S_h(p,\Theta_N) = QD_h, \quad (4.3)$$

Note that the aggregate offer curve is the sum of the Ideal Offer Curves evaluated at Θ_n^{NC} (n=1,2,..,5) for the five strategic suppliers and Θ_n (n=6,…,N) for remaining suppliers. This price is lower than PC_h^F because it assumes that no quantity offers are prevented from selling energy because of the transmission constraints. For this reason, it provides an lower bound on the market-clearing price that would result if all strategic suppliers faced the Ideal Residual Demand curve instead of the Feasible Residual Demand curve but kept the same fixed-price forward contract obligations.

As noted earlier, if a supplier faces greater competition during all hours of the year because that supplier does not expect quantity offers from other suppliers to be preventing from selling energy because of transmission constraints, that supplier is potentially more likely to sell more fixed-price forward contract obligations in order to pre-commit to being a more aggressive competitor (submit offer curves closer to its marginal cost curve) in the short-term market. Neither of the two counterfactual prices attempts to capture this additional source of potential competitiveness benefits from a commitment to transmission investments that significantly reduce the frequency and magnitude of transmission congestion.

Following the analogous logic to computing the Predicted Feasible Actual market-clearing price, a Predicted Ideal Actual market-clearing price can be computed by constructing an aggregate supply curve from the sum of the Ideal Offer Curve for all suppliers. Mathematically, the Predicted Ideal Actual market-clearing price, PP_h^I, is the smallest price that solves:

$$S_h(p,\Theta_1) + S_h(p,\Theta_2) + , , , + S_h(p,\Theta_5) + S_h(p,\Theta_6), + \ldots, S_h(p,\Theta_N) = QD_h, \quad (4.4)$$

This price should be less than or equal to the actual market-clearing price because it assumes that the Ideal Offer Curves are used for all suppliers, including the five strategic suppliers. Particularly, during the high-priced hours of the day, PP_h^I is significantly less than the actual market-clearing price and the Predicted Feasible Actual market-clearing price.

Figure 9 plots the daily average actual price and the daily average Predicted Ideal Actual price. Although the daily average Predicted Ideal Actual prices follow the same general pattern as the daily average actual prices, they are typically lower and less volatile than the actual prices. This result suggests that even without a change in suppler offer behavior, increasing the amount

of transmission capacity to reduce the number and total volume of offer quantities that cannot sell energy because of transmission constraints has significant consumer benefits in terms of lower average wholesale prices and less volatile wholesale prices.

For each of the two counterfactual prices, I compute two measures of the competitiveness benefits of transmission investments that commit to a reduced frequency of congestion. The first is the difference between the actual market price and the counterfactual price times the total demand in the AESO. The second is a relative measure, the reduction in wholesale market costs as a percentage actual wholesale market costs, the actual market-clearing price times the total demand in the AESO. In terms of our previously defined notation, the first two hourly measures are:

$$\Delta R_h^F = (P_h - PC_h^F)QD_h \text{ and } \Delta R_h^I = (P_h - PC_h^I)QD_h, \quad (4.5)$$

which are the difference in wholesale market costs from consumers paying the counterfactual Feasible Market Price and the difference in wholesale market costs from consumers paying the counterfactual Ideal Market Price. The second two measures are the ratio of the difference in wholesale market cost over some time horizon divided by actual wholesale market costs over that same time horizon. Let H equal the number of hours in that time horizon, then

$$\Delta RR_h^F = 100 * \frac{\sum_{h=1}^{H}(P_h - PC_h^F)QD_h}{\sum_{h=1}^{H} P_h * QD_h} \text{ and } \Delta RR_h^I = 100 * \frac{\sum_{h=1}^{H}(P_h - PC_h^I)QD_h}{\sum_{h=1}^{H} P_h * QD_h}, \quad (4.4)$$

which are the change in wholesale energy costs over horizon H as a percent of actual wholesale energy purchase costs over horizon H for both the Feasible and Ideal counterfactual prices.

Table 6 lists the annual average of the hourly wholesale cost changes for the Ideal and Feasible Counterfactual Prices for 2009, 2010 and 2011. It also lists the average hourly wholesale cost changes for the entire sample period. Third column of the table lists the average hourly wholesale market revenue for each year and for the entire sample. The fifth column shows the annual average hourly wholesale cost difference using the Ideal Counterfactual price as a percentage of annual average hourly wholesale market revenues. The last row in the table gives the sample hourly average hourly wholesale cost difference using the Ideal Counterfactual price as a percentage of sample average of hourly wholesale market revenues. The last column shows the annual average hourly wholesale cost difference using the Feasible Counterfactual Price as a percentage of annual average hourly wholesale market revenues. The last row gives

the sample average hourly wholesale cost difference using the Feasible Counterfactual Price as a percentage of sample average of hourly wholesale market revenues.

Table 6: Annual and Sample Average Hourly Revenue Differences for Ideal and Feasible Counterfactual Prices in CAD and as Percentage of Annual Wholesale Energy Costs

Year	Ideal Price Cost Difference	Feasible Price Cost Difference	Wholesale Energy Costs	Ideal Cost Difference as a Percent of Wholesale Revenues	Feasible Price Cost Difference as a Percent of Wholesale Revenues
2009	61,912.99	2,734.43	398,345.3	15.54254	0.686447
2010	81,648.03	2,080.56	426,525.7	19.14258	0.487792
2011	102,963.6	5,043.68	653,753	15.74962	0.771496
Sample	79,590.19	3,066.67	472,816.4	16.83321	0.648596

Figure 10(a) and 10(b) plot the monthly average values of the hourly wholesale cost changes for the Feasible and Ideal Counterfactual Prices. The average monthly demand served in the AESO is also plotted in each figure. The average monthly wholesale cost changes using the Feasible Counterfactual Price shown in Figure 10(a) finds modest, but economically significant competitiveness benefits from suppliers submitting offer prices under the expectation of no congestion, but actually facing the same amount of congestion as actually occurred during that hour. Although the average hourly revenue change over the sample is 3,067 CAD, during one month it exceeded 25,000 CAD. Comparing the pattern of the monthly average demand in the AWEM to the monthly average values of the Feasible Counterfactual Price wholesale cost difference shows a positive correlation between the two monthly values.

Figure 10(b) finds substantially larger revenue changes associated with the strategic suppliers submitting offer prices under the expectation of no congestion and the realization that there is actually no congestion, the Ideal Counterfactual Price hourly wholesale cost difference. The sample average hourly wholesale cost difference using the Ideal Counterfactual Price is 79,590 CAD. There is even a month when the average hourly wholesale cost difference with the Ideal Counterfactual price is greater than 500,000 CAD. There appears to be a positive correlation between the monthly average value of this cost difference and the monthly average value of demand in the AWEM.

The pattern of the monthly value of the wholesale cost differences using in the Feasible Counterfactual price as a percentage of actual monthly wholesale market revenues in Figure

11(a) replicates the pattern of the monthly wholesale cost differences in Figure 10(a). For the entire sample the Feasible Counterfactual price wholesale cost difference is 0.64 percent of total wholesale energy costs. However, during certain months, this percentage is substantially higher. In fact, it is more than 2 percent of monthly wholesale energy costs during one month of the sample.

For the entire sample, the Ideal Counterfactual price wholesale cost difference is 16.8 percent of total wholesale energy costs. As shown in Figure 11(b), during certain months, this percentage is substantially higher, and in one month more than 45 percent of actual wholesale market revenues. Although for most of the months this percentage is below 20 percent, it never fall below 5 percent, indicating that during all months of the sample period there are substantial competitiveness benefits from suppliers expecting there to be no transmission constraints that prevent quantity increments offered by them and their competitors from selling energy and this expectation in fact turns out to be case.

5. Conclusions

These empirical results demonstrate economically sizeable competitiveness benefits from facing strategic suppliers with residual demand curves that reflect little likelihood that transmission constraints will limit the quantity increments of other firms from selling energy. Even if these expectations do not turn out to be the case, because strategic suppliers with these expectations about the extent of competition that they face are predicted to submit lower offer prices, the resulting market-clearing prices, even with the same amount of transmission congestion as actually occurred, will be lower. These Feasible Counterfactual Offer Curve market-clearing prices imply sizeable average wholesale cost differences, an average of 3,067 CAD per hour. Over the three–year sample, the total wholesale cost difference from the five largest strategic suppliers in AWEM expecting that none of the quantity increments of their competitors will be unable to supply energy because of transmission constraints is more than 94 million CAD, even if there were no change in the actual realized transmission congestion.

If these expectations of limited congestion by the strategic suppliers actually hold and no suppliers are actually prevented from selling energy because of transmission constraints and the Ideal Counterfactual Offer market-clearing prices are the relevant price paid by electricity consumers, the total wholesale cost savings for the sample period is more than $2 billion dollars.

Clearly, this amount of wholesale cost savings over a three-year period could fund a substantial amount of transmission expansions.

Taken together, these results provide persuasive empirical evidence that the competitiveness benefits of transmission expansions should be accounted for in the transmission planning processes for formal wholesale electricity markets. Given the magnitude of these benefits, many transmission expansions with net economics benefits to electricity consumers may not be undertaken because this source of economic benefits is not accounted for. This is particularly the case for the AWEM market given the ownership shares of generation capacity of the five strategic suppliers and the dominant share that coal and natural gas-fired generation plays in the electricity supply mix. The extremely steep offer curves that suppliers submit, particularly during periods when there is likely to be transmission congestion, argues in favor of a transmission policy that accounts for these competitiveness benefits.

These results also support the view that planning and constructing the transmission network in Alberta in a forward-looking manner to limit the frequency and magnitude of congestion can yield sizeable net benefits to electricity consumers in the province as demonstrated by both the Feasible and Ideal Counterfactual price wholesale market cost differences changes.

Finally, it is important to emphasize that a potentially sizeable source of additional competitiveness benefits was not accounted for in this analysis. Specifically, the incentive for a supplier to change its fixed-price forward contract obligations in response to the reduced number of opportunities to exercise unilateral market power because of the increased competition it faces because of the significantly reduced frequency and magnitude of transmission congestion is not accounted for. Such an analysis would require information on the fixed-price forward market obligations of the five largest strategic suppliers in the AWEM. This data is currently considered confidential by market participants and is not available to the AESO. However, given the current concentration of generation ownership in the AWEM and the structure of offer curves submitted to the AESO during the sample period, this forward contracting competitiveness benefit from a transmission planning and construction policy that limits the frequency and magnitude of transmission congestion is likely to be economically significant.

References

Arellano, M. Soledad, and Serra, Pablo (2008) "The Competitive Role of the Transmission System in Price-Regulated Power Industries," *Energy Economics*, vol. 30, 1568-1576.

Borenstein, Severin; Bushnell, James, B. and Stoft, Steven (2000) "The Competitive Effects of Transmission Capacity in a Deregulated Electricity Industry," *RAND Journal of Economics*, vol. 31, No. 2, 294-325.

McRae, Shaun D. and Wolak, Frank A. (2012) "How Do Firms Exercise Unilateral Market Power? Evidence from a Bid-Based Wholesale Electricity Market," in Eric Brousseau and Jean-Michel Glachant (editors), Manufacturing Markets: legal, political and economic dynamics" Cambridge University Press.

Wolak, Frank A. (2000), "An Empirical Analysis of the Impact of Hedge Contracts on Bidding Behavior in a Competitive Electricity Market," *International Economic Journal*, 14(2), 1-40.

Wolak, Frank A. (2003), "Identification and Estimation of Cost Functions Using Observed Bid Data: An Application to Electricity Markets," in M. Dewatripont, L.P. Hansen, and S.J. Turnovsky, eds., *Advances in Economics and Econometrics: Theory and Applications, Eight World Congress, Volume II*. New York: Cambridge University Press, pp. 133-169.

Wolak, Frank A. (2007), "Quantifying the Supply-Side Benefits from Forward Contracting in Wholesale Electricity Markets," *Journal of Applied Econometrics*, 22, 1179-1209.

Figure 1(a): Annual Demand Duration Curves for 2009, 2010, and

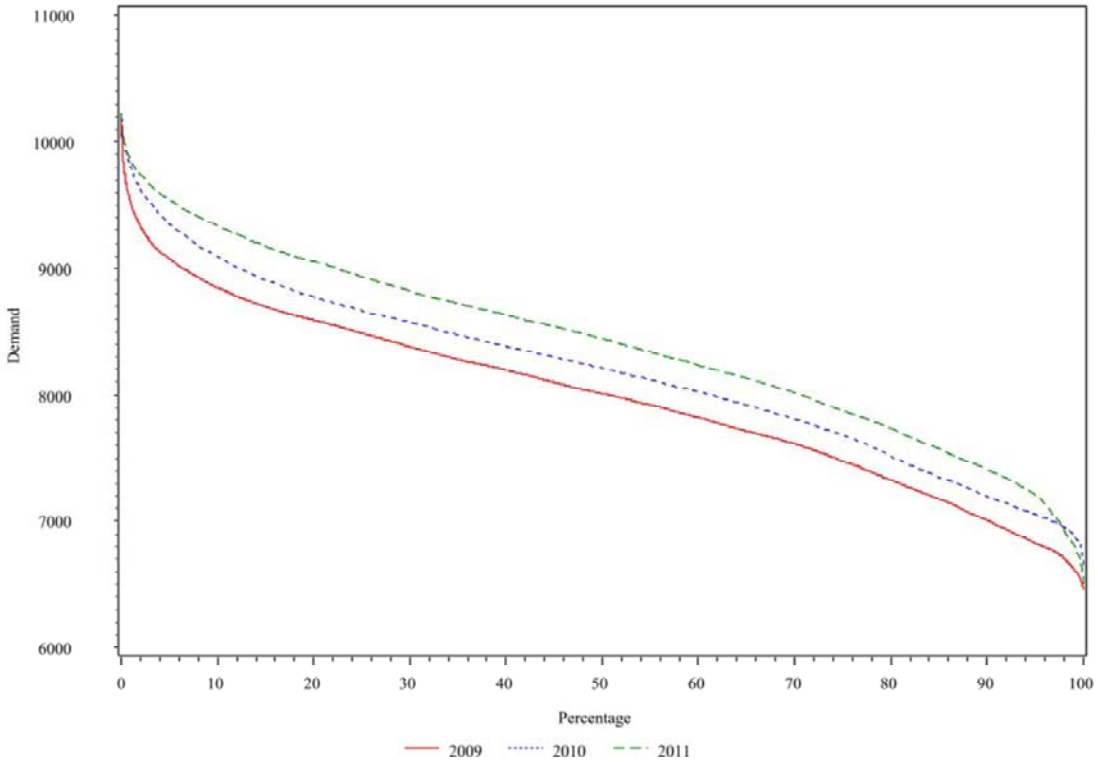


Figure 1(b): Highest 1 Percent of Annual Demand Duration Curves 2009, 2010,

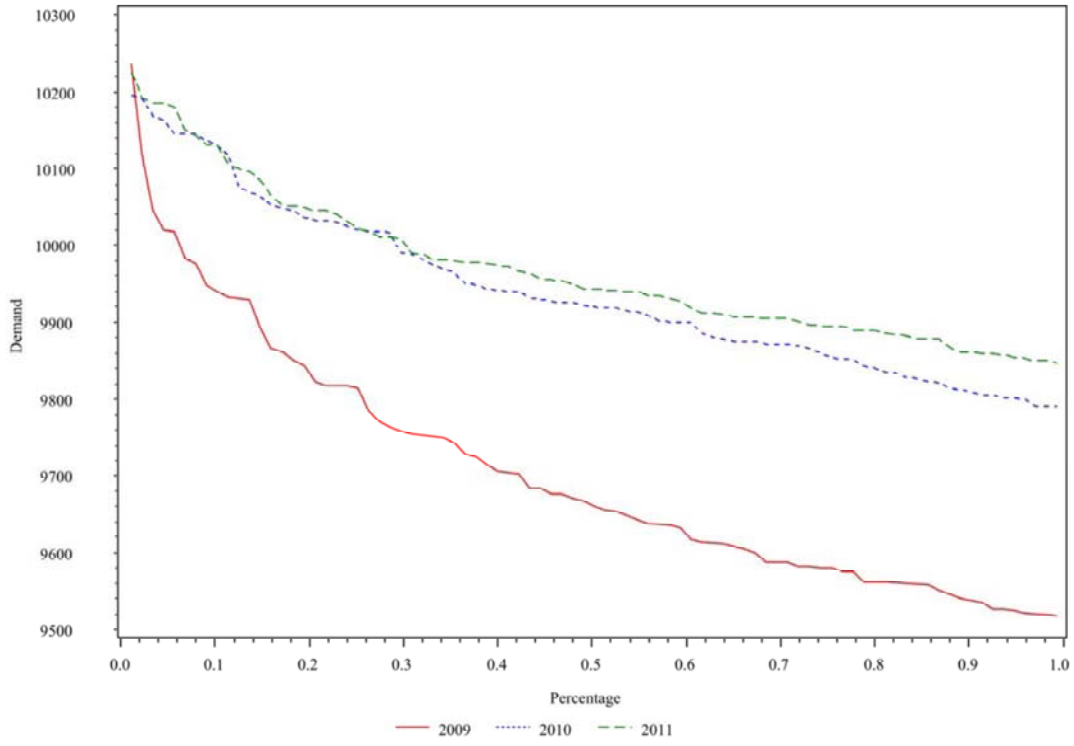


Figure 2(a): Annual Price Duration Curves 2009, 2010, and 2011

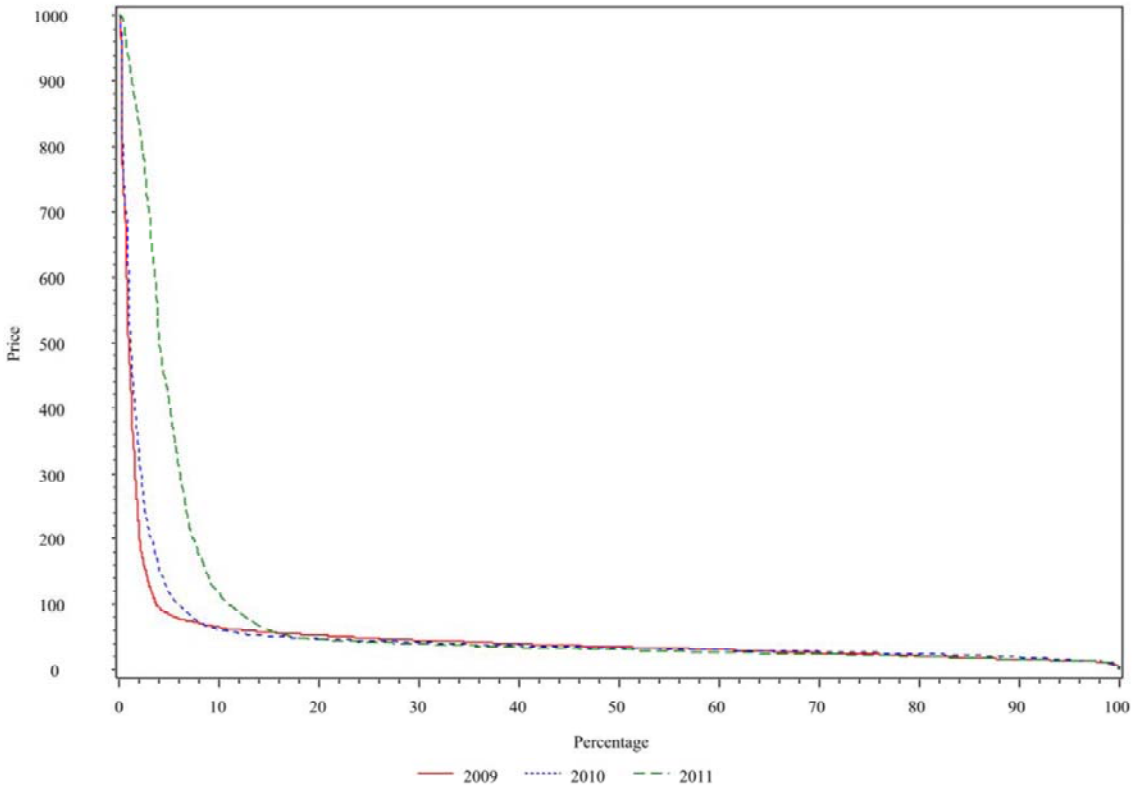


Figure 2(b): Upper 10 Percent of Annual Price Duration Curves 2009, 2010, 2011

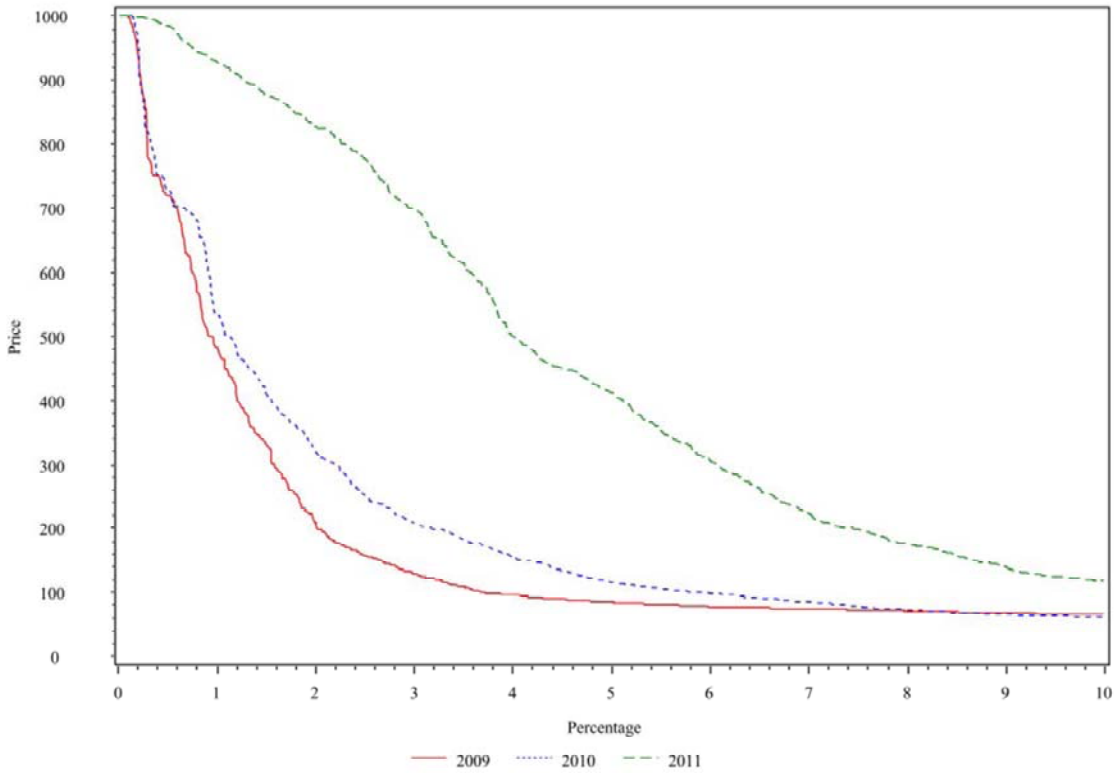


Figure 3: Ideal and Feasible Aggregate Offer Curve for Hour 12 of 5/12//2010

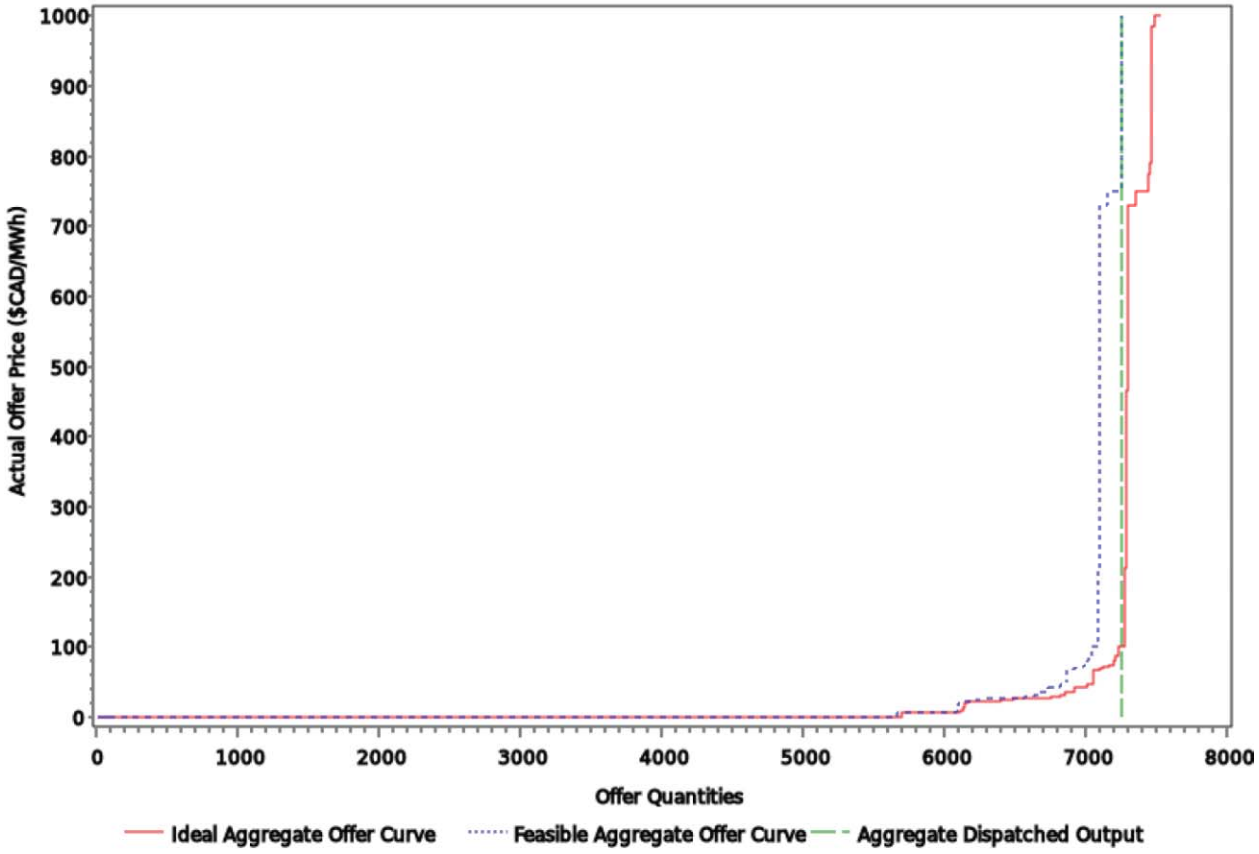


Figure 4(a): Ideal and Feasible Residual Demand Curves for ATCO Power, Hour 13 of 5/16/2010

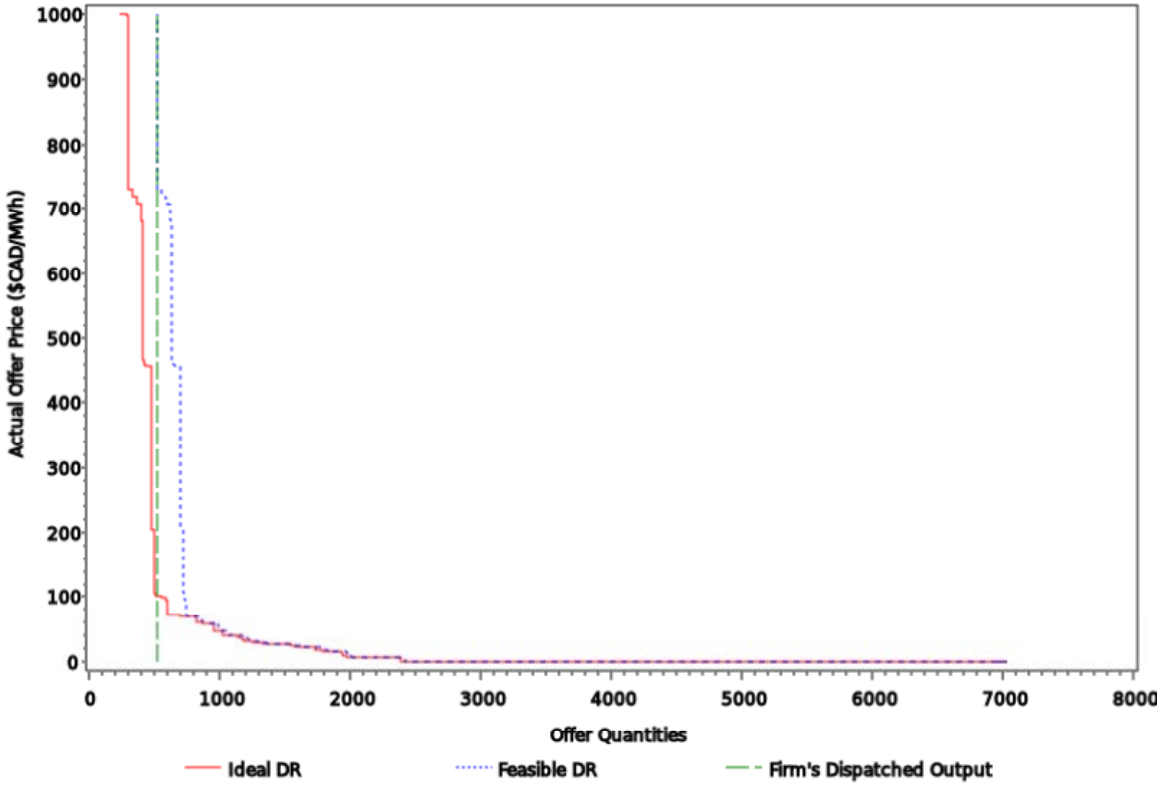


Figure 4(b): Ideal and Feasible Residual Demand Curves for Capital Power, Hour 13 of 5/16/2010

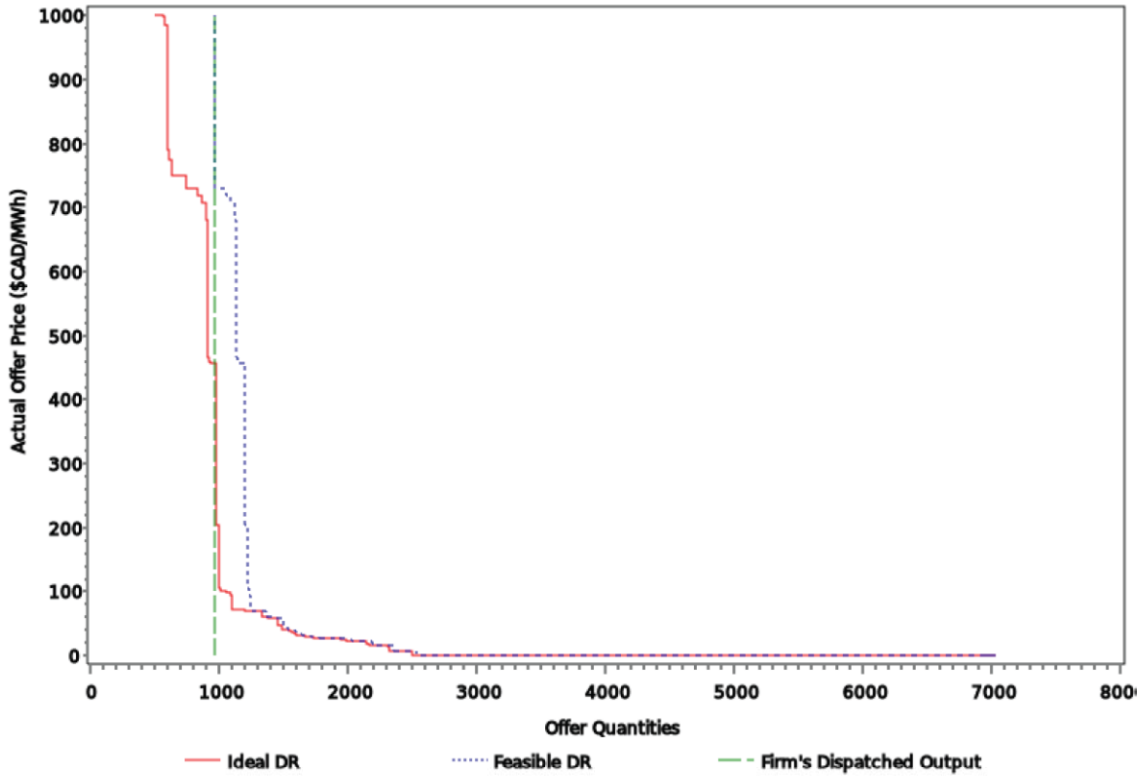


Figure 4(c): Ideal and Feasible Residual Demand Curves for TransAlta, Hour 13 of 5/16/2010

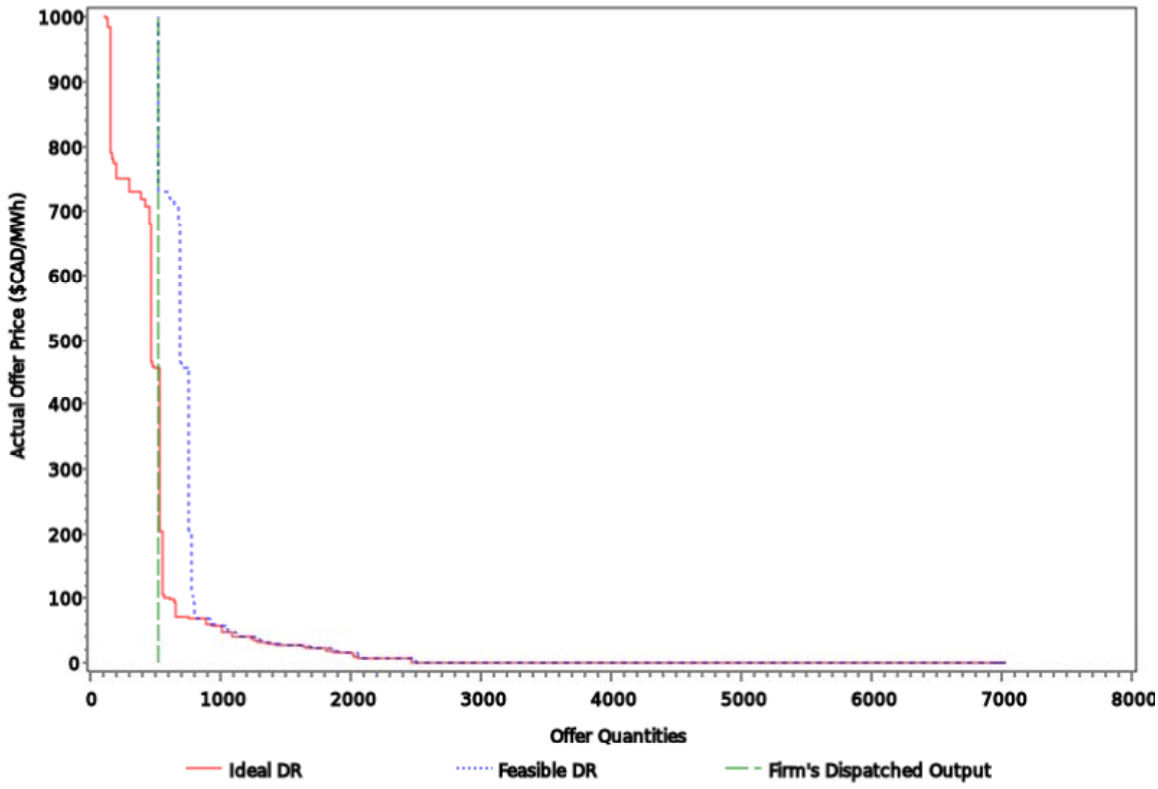


Figure 4(d): Ideal and Feasible Residual Demand Curves for ENMAX, Hour 13 of 5/16/2010

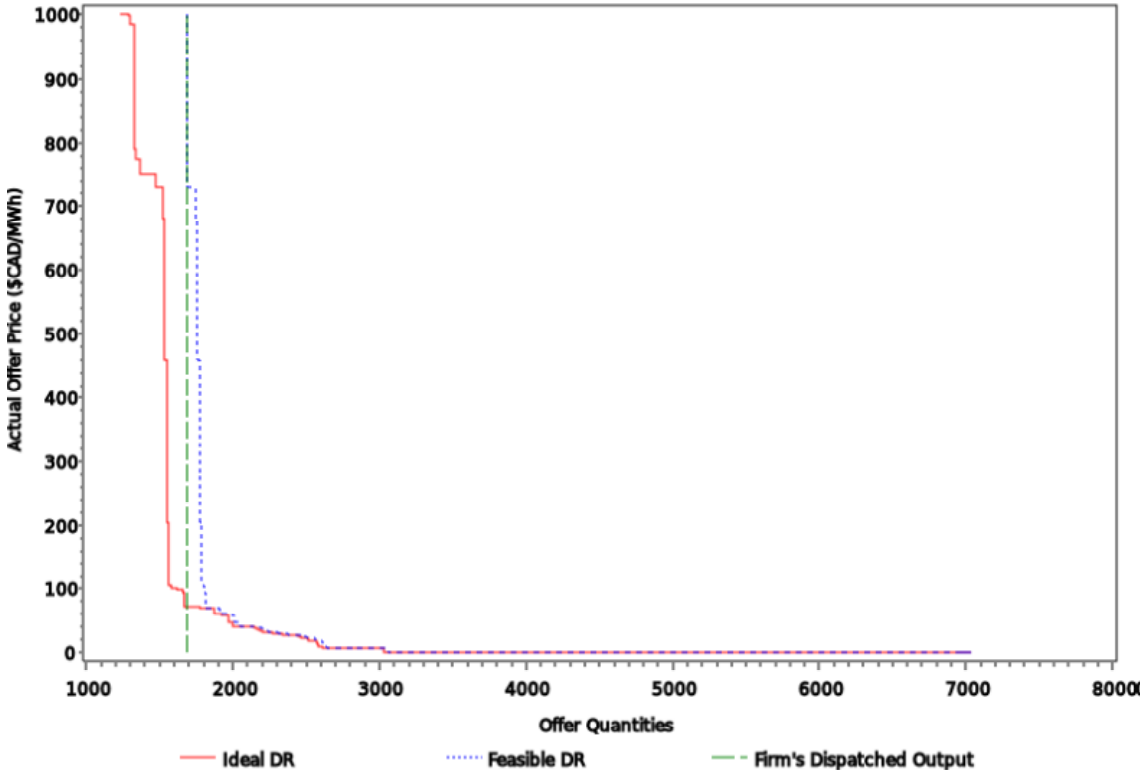


Figure 4(e): Ideal and Feasible Residual Demand Curves for TransCanada, Hour 13 of 5/16/2010

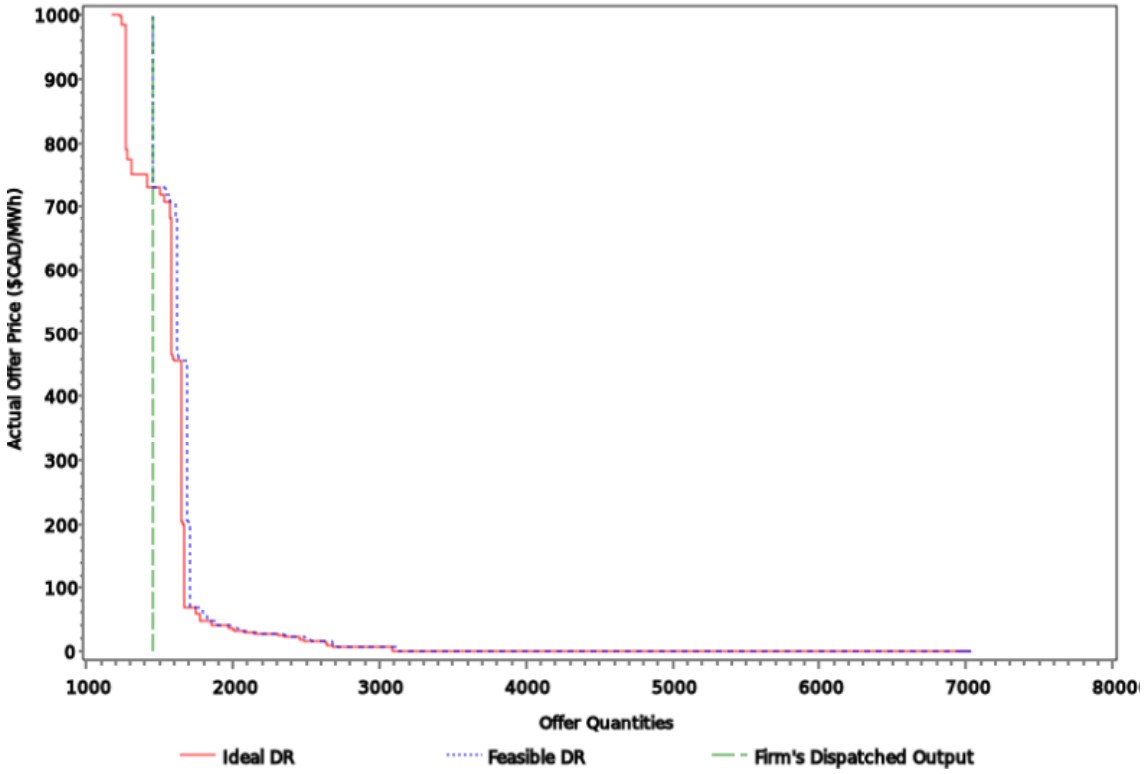


Figure 5(a): Derivation of Expected Profit-Maximizing Offer Curve

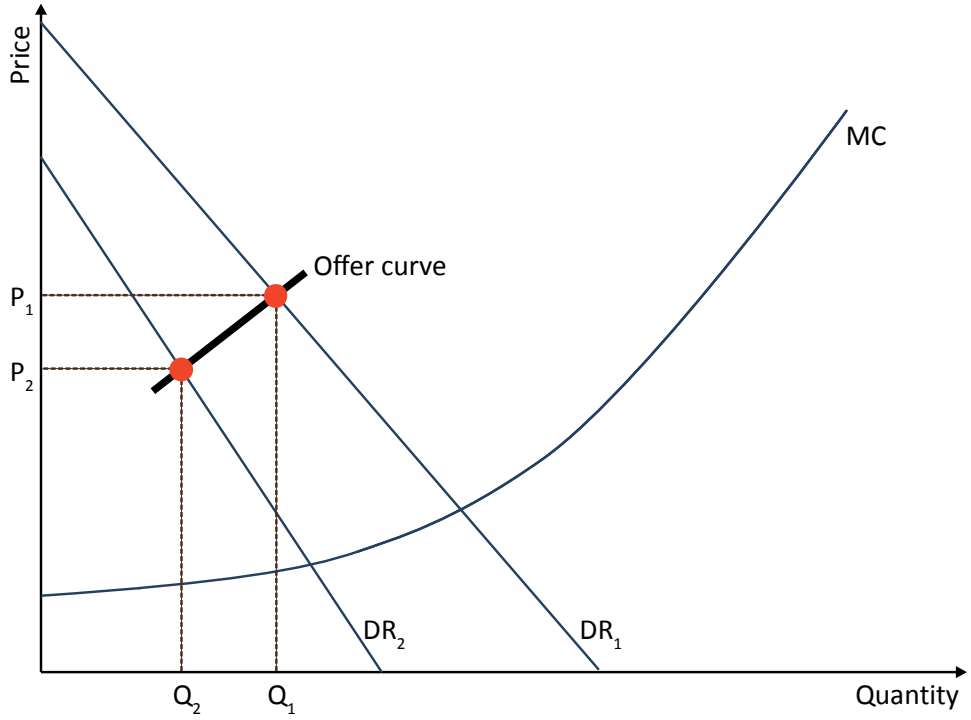


Figure 5(b): Expected Profit-Maximizing Offer Curve (flatter residual demands)

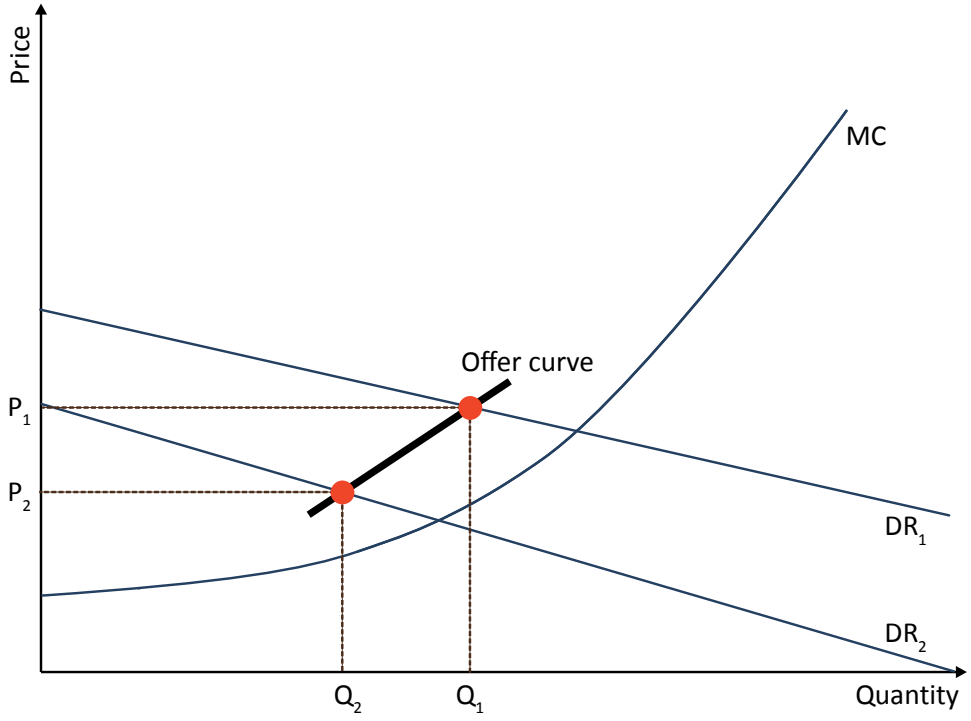


Figure 5(c): Expected Profit-Maximizing Offer Curve (perfectly elastic residual demands)

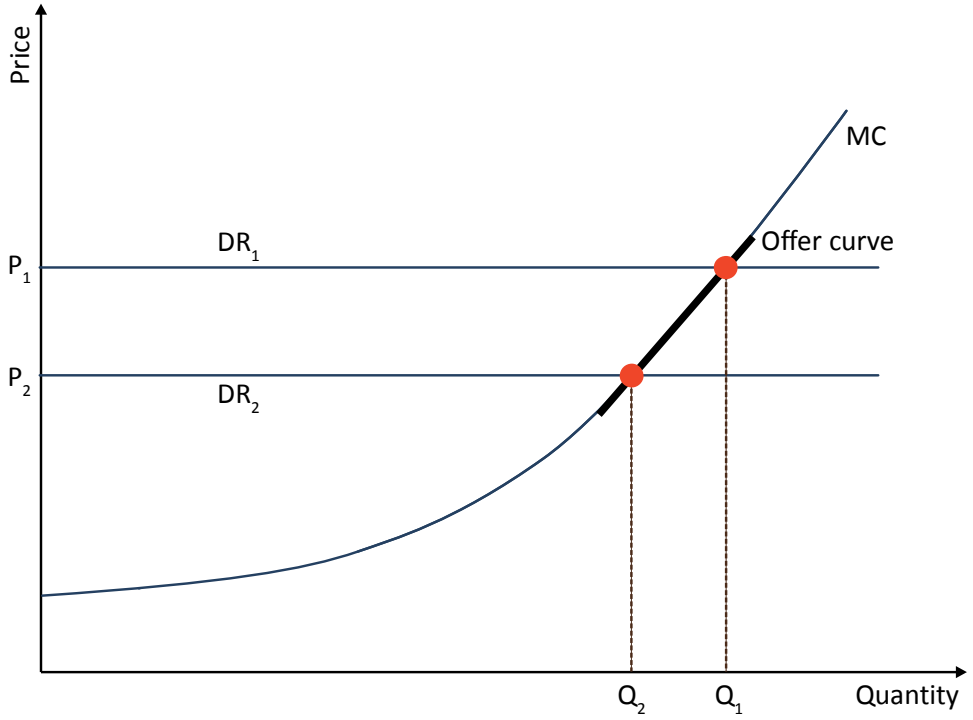


Figure 5(d): Impact of Step Functions on Expected Profit-Maximizing Offer Curve

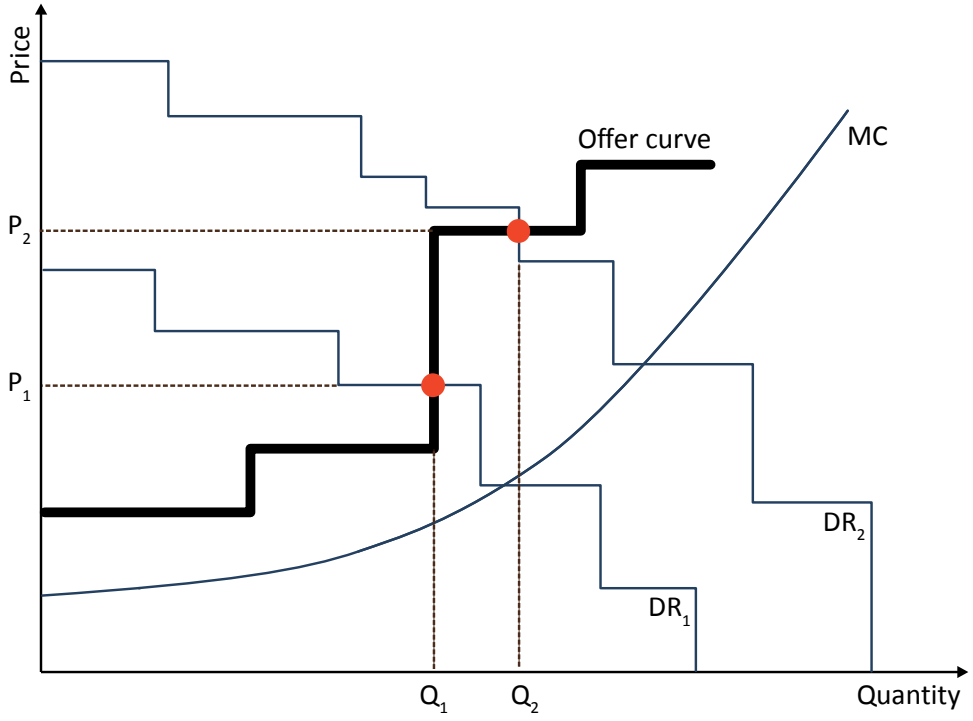


Figure 5(e): Expected Profit-Maximizing Step-Function Offer Curve

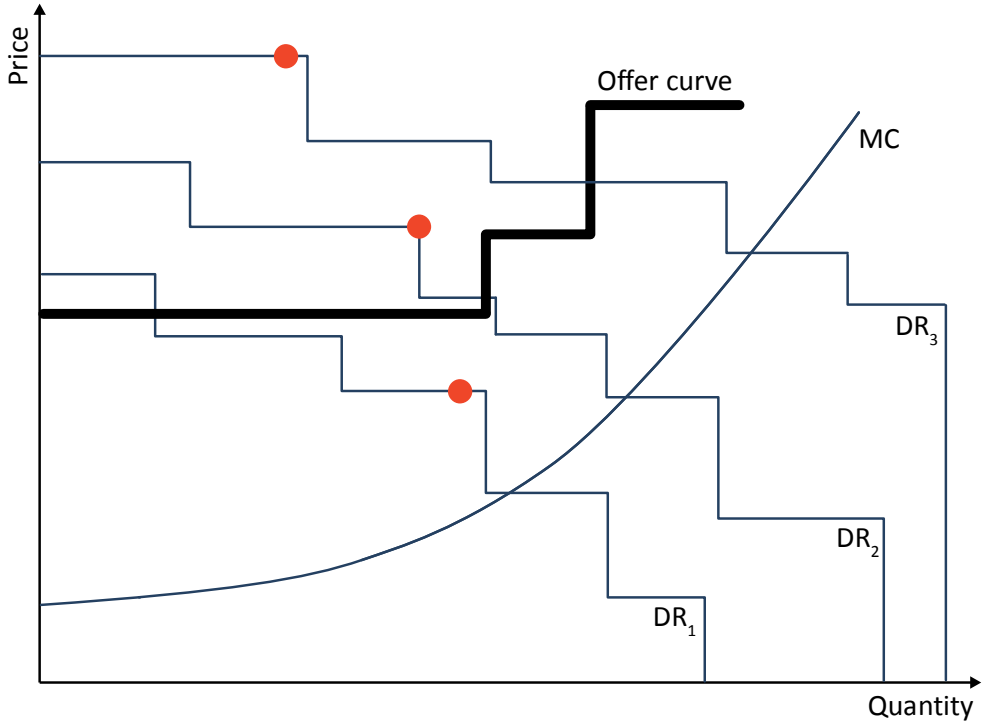


Figure 6: Sample Calculation of Hourly Offer Price for Firm A

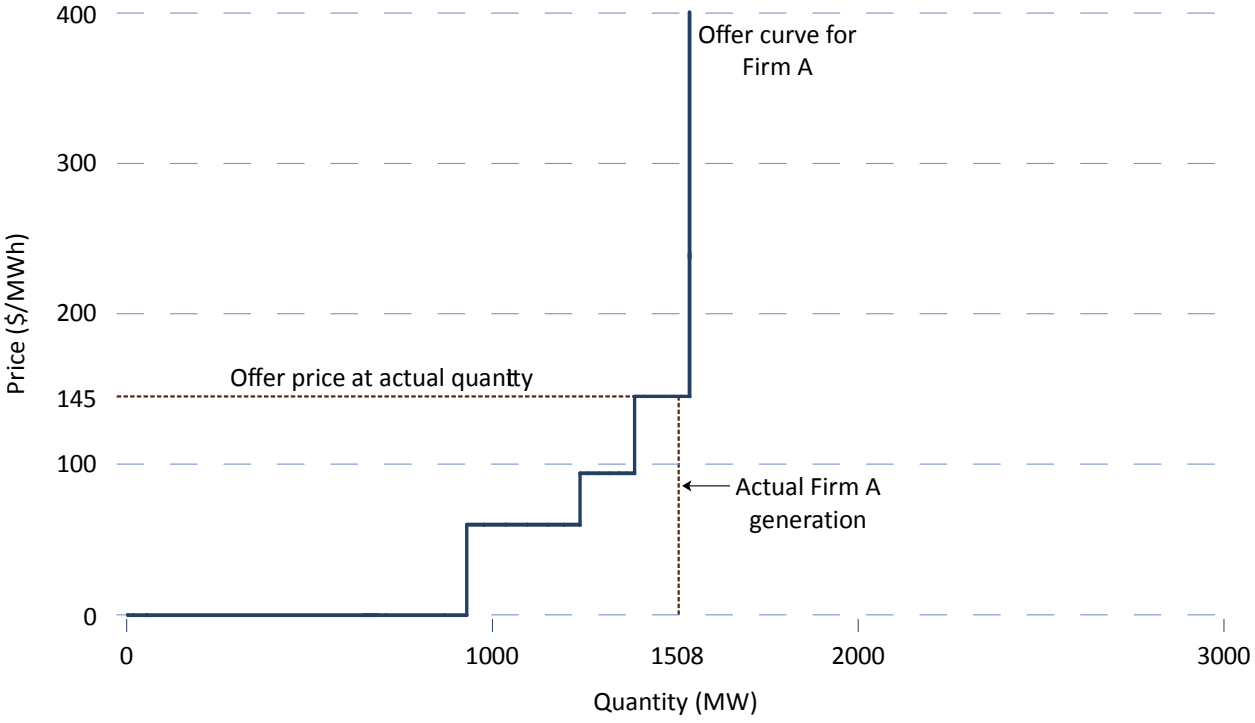


Figure 7(a): Actual and Shifted No-Congestion Offer Curves for Firm A

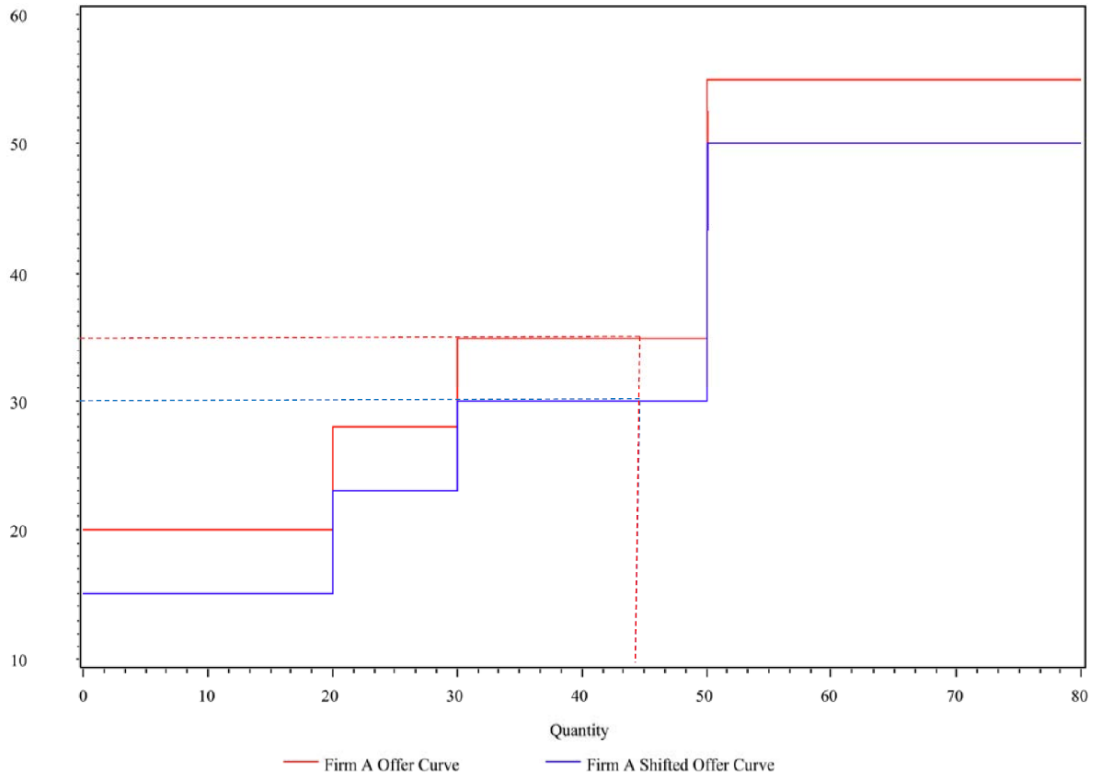


Figure 7(b): Actual and Shifted No-Congestion Offer Curves for Firm B

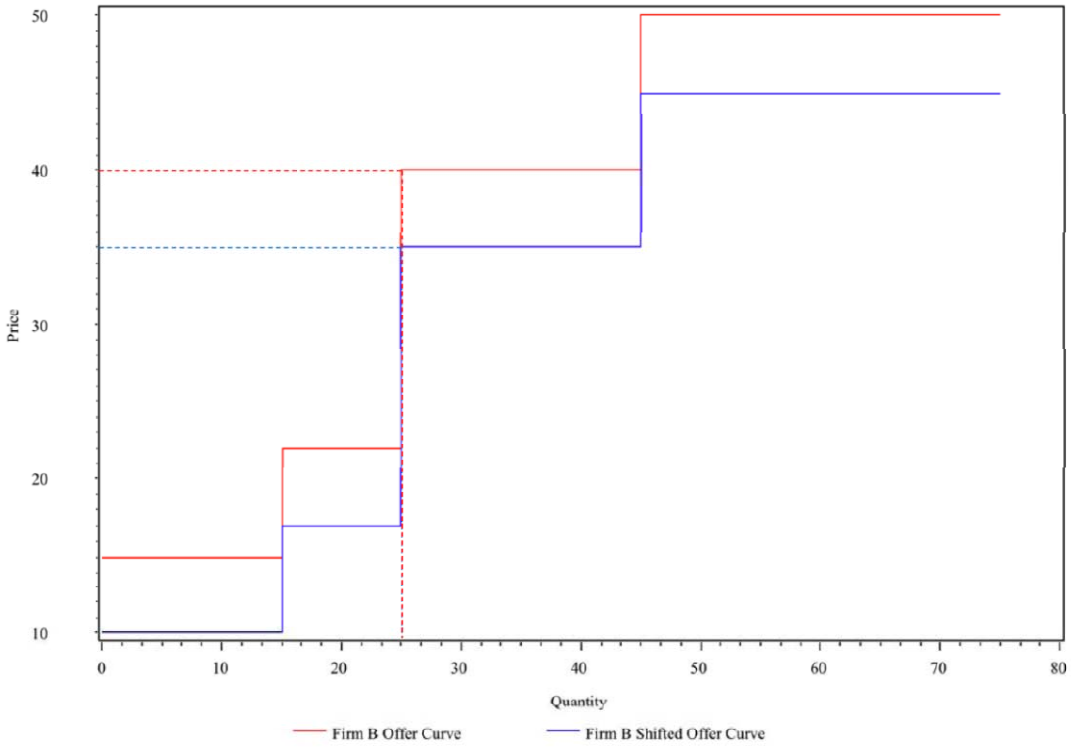


Figure 7(c): Actual and Shifted No-Congestion Aggregate Offer Curves for Firm A and Firm B

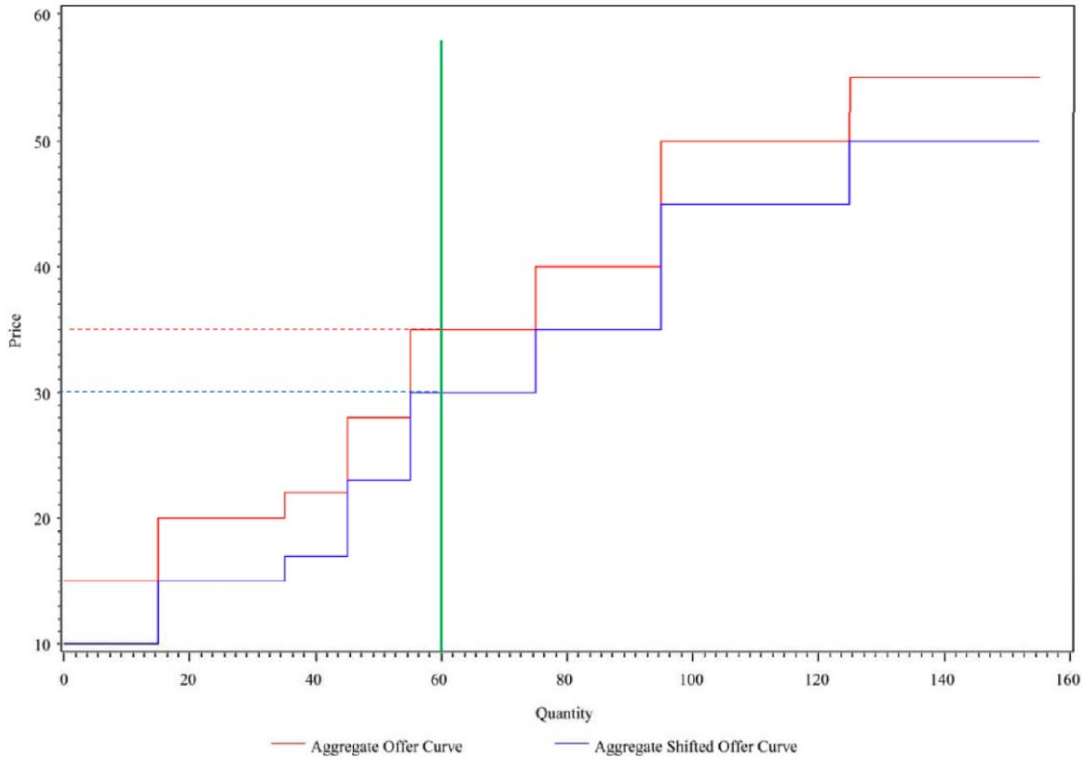


Figure 8: Daily Average Actual Prices and Predicted Feasible Actual Prices

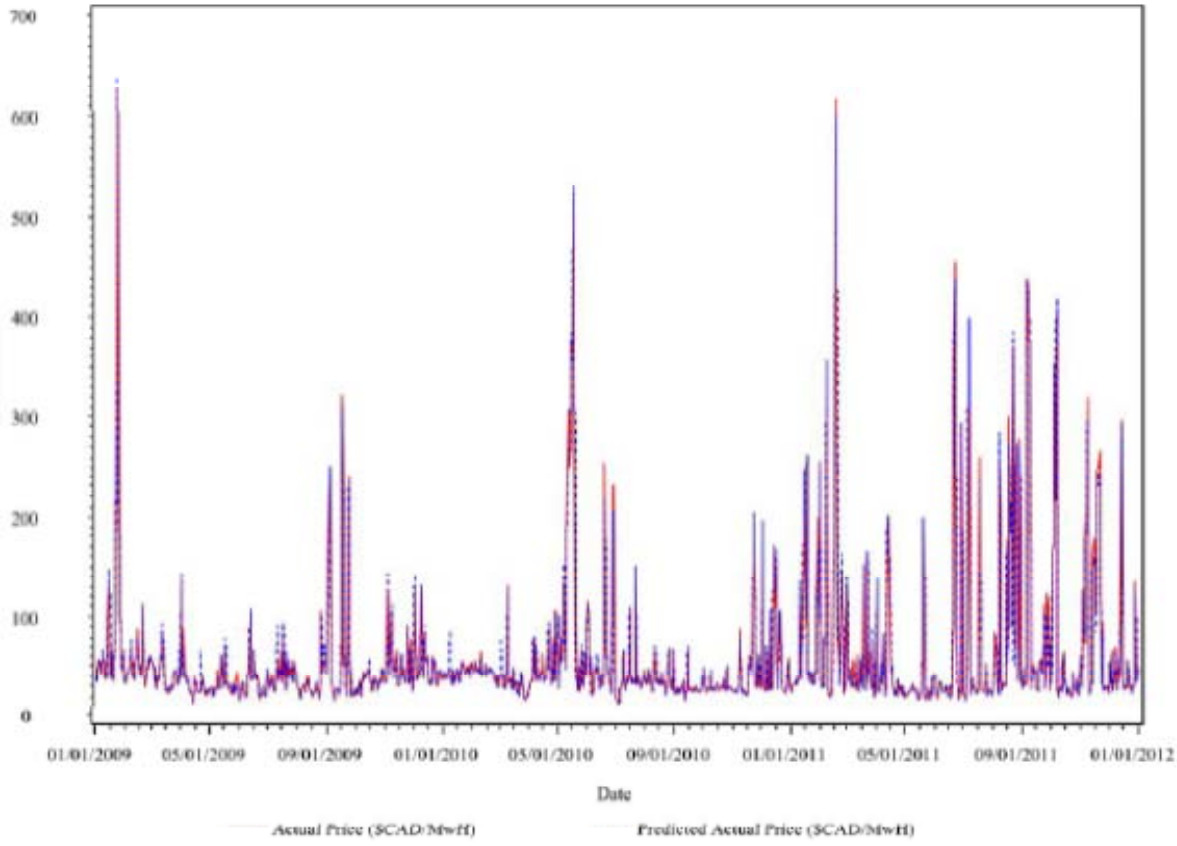


Figure 9: Daily Average Actual Prices and Predicted Ideal Actual Prices

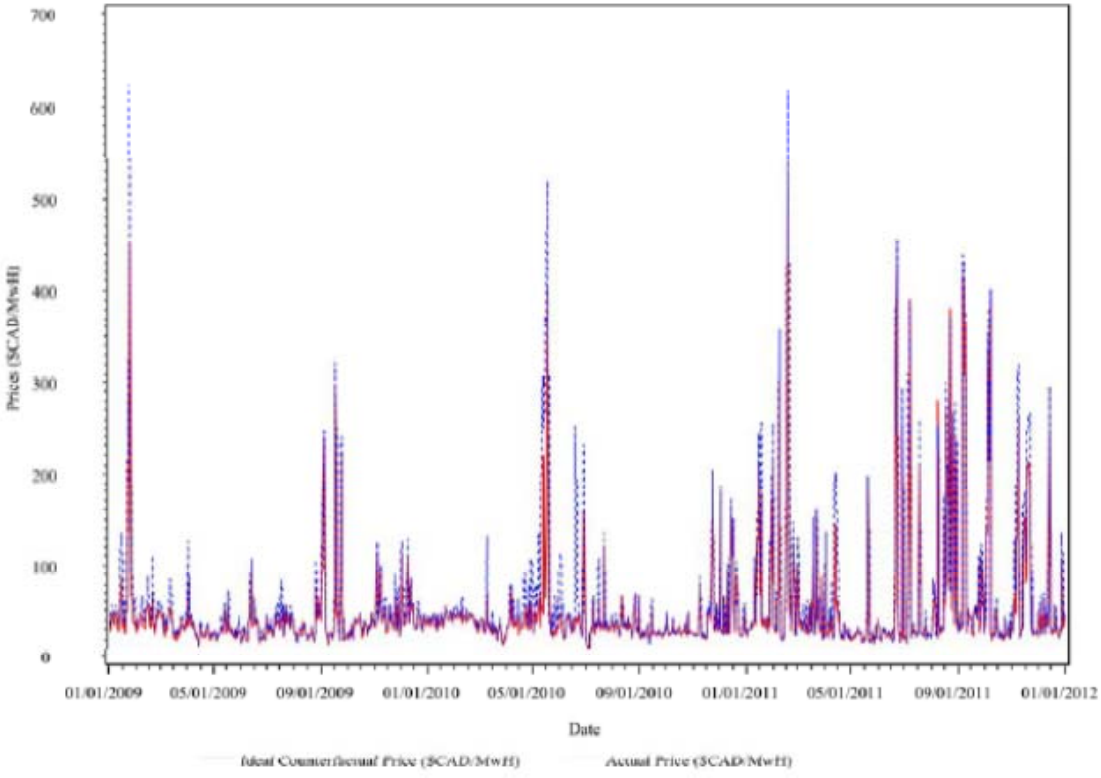


Figure 10(a): Monthly Average Wholesale Revenue Change with Feasible Price and Monthly Average Demand

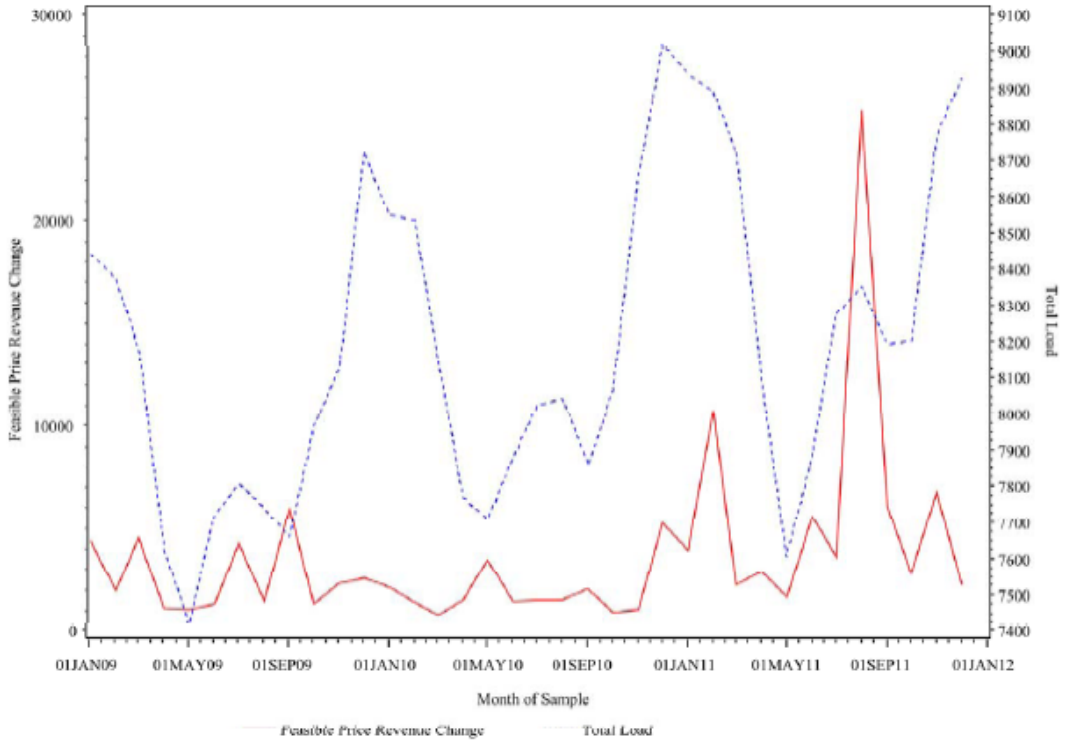


Figure 10(b): Monthly Average Wholesale Revenue Change with Ideal Price and Monthly Average Demand

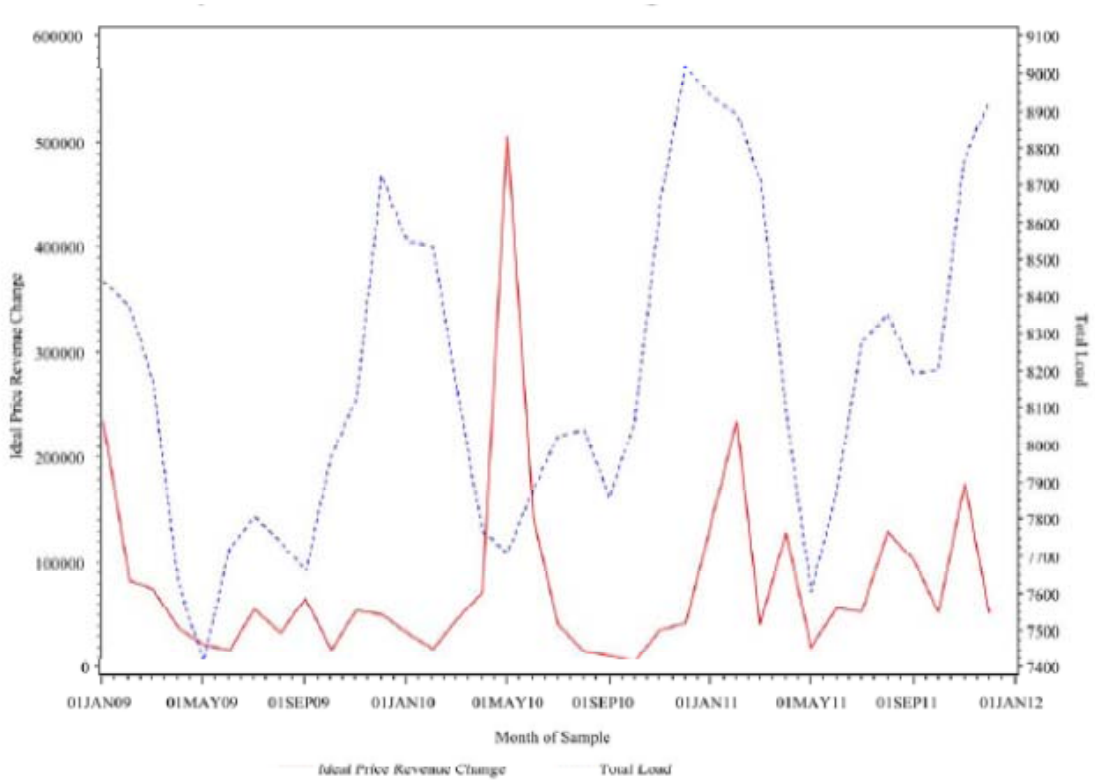


Figure 11(a): Monthly Wholesale Revenue Change with Feasible Price as a Percentage of Actual Monthly Wholesale Revenues and Monthly Average Demand

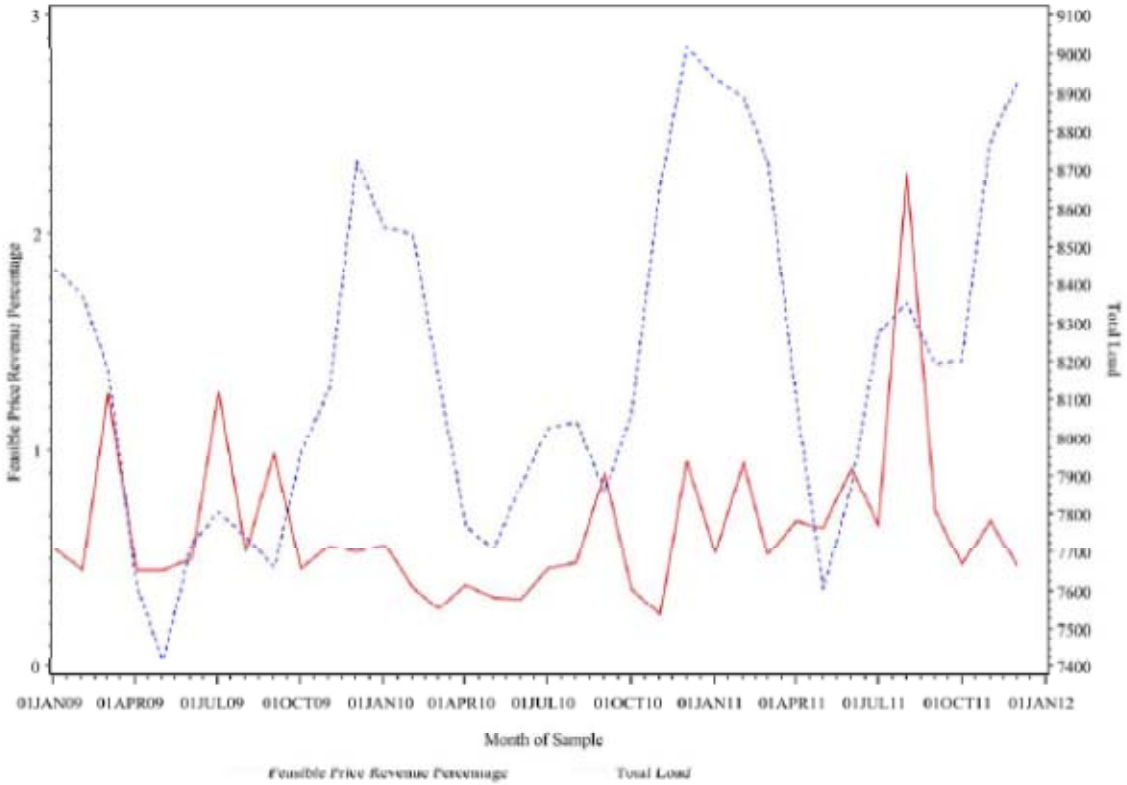
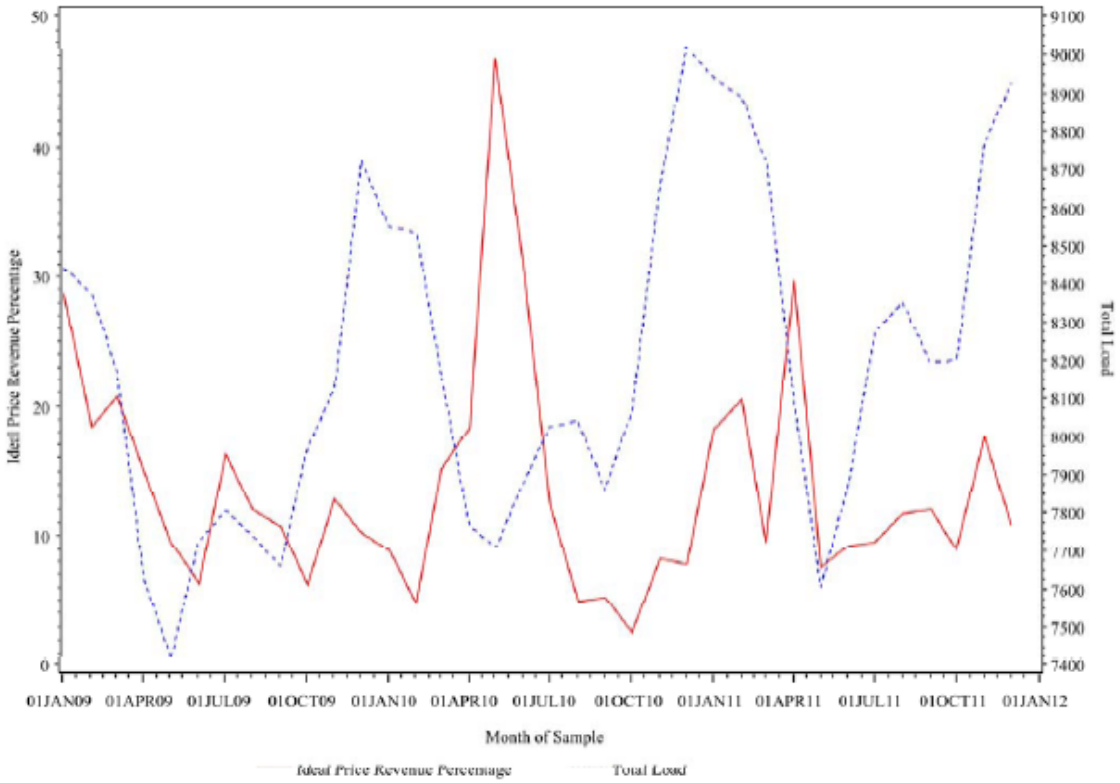


Figure 11(b): Monthly Wholesale Revenue Change with Ideal Price as a Percentage of Actual Monthly Wholesale Revenues and Monthly Average Demand



Measuring the Performance of Large-Scale Combinatorial Auctions: A Structural Estimation Approach *

Sang Won Kim †, Marcelo Olivares ‡, and Gabriel Y. Weintraub §

May 2, 2012

Abstract

The main advantage of a procurement combinatorial auction (CA) is that it allows suppliers to express cost synergies through package bids. However, bidders can also strategically take advantage of this flexibility, reducing the performance of the auction. In this paper, we develop a structural estimation approach for large-scale first-price CAs. We use bidding data to estimate the firms' cost structure and evaluate the performance of the auction in terms of the cost efficiency of the allocation and payments to the bidders. To overcome the computational difficulties arising from the large number of bids observed in large-scale CAs, we propose a novel simplified model of bidders' behavior where markups of each package bid are chosen based on a reduce set of package characteristics. We apply our method to the Chilean school meals auction, in which the government procures half a billion dollar worth of meal services every year and bidders submit thousands of package bids. Our estimates suggest that bidders' cost synergies are economically significant in this application, and the current CA mechanism achieves high allocative efficiency and a reasonable procurement cost. We also perform counterfactuals to compare the performance of the current CA with alternative mechanisms such as VCG.

Keywords: combinatorial auctions, procurement, empirical, structural estimation, auction design, public sector applications.

*Acknowledgements: We thank Rafael Epstein for the valuable insights he provided to us. We are thankful for the helpful comments received from Lanier Benkard, Estelle Cantillon, Awi Federgruen, Jeremy Fox, Ali Hortaçsu, Jakub Kastl, Paul Milgrom, Martin Pesendorfer, Mar Reguant, Michael Riordan, Richard Steinberg, Assaf Zeevi, and seminar participants at Caltech SISL, Columbia Business School, Chicago Booth, Duke Fuqua, Informs Conference, London School of Economics, MSOM Conference, Revenue Management Conference, Stanford MS&E, Stanford Structural IO Lunch, and Utah Winter Operations Conference. We thank Daniel Yung for exemplary research assistance. We also thank the Social Enterprise Program, CIBER, and the W. Edwards Deming Center at Columbia Business School for their financial support.

†Columbia Business School, Email: skim14@gsb.columbia.edu,

‡Columbia Business School, Email: molivares@columbia.edu

§Columbia Business School, Email: gweintraub@columbia.edu

1 Introduction

In many important procurement settings suppliers face cost synergies driven by economies of scale or density. For example, transportation service providers can lower costs by coordinating multiple deliveries in the same route, and producers can lower average costs by spreading a fixed cost across several units. Motivated by this type of settings, auction mechanisms that allow bidders to submit package bids for multiple units so that they can express their synergies have received much recent attention in practice and theory. In fact, these multi-unit auctions have been successfully implemented in many applications, including the procurement of school meal services, bus routes, electricity, transportation services, and inputs in private firms, as well as in non-procurement settings, such as the auctions for wireless spectrum run by the Federal Communications Commission (FCC).

In this paper we introduce a structural estimation approach to empirically analyze the performance of first-price combinatorial auctions (CAs), a multi-unit mechanism that allows bidders to submit separate bids for different combinations or packages of units (see Cramton et al. (2006) for an overview on CAs). We address a central design question: what is the impact that allowing package bidding via a CA has on performance. Our analysis considers two performance measures: (1) efficiency, which relates the actual *bidders' costs* incurred in the CA allocation relative to the minimum possible cost allocation any subset of suppliers could achieve; and (2) optimality, which relates to the total expected payments to bidders by the auctioneer.

There are two countering effects that can affect the performance of a CA. The main advantage of package bidding is that it allows bidders to express cost synergies in their bids. In contrast, if bidders were allowed only to submit bids for each unit separately they would face the risk of winning some units but not others. This phenomenon, known as the *exposure problem*, makes the bidders less aggressive in expressing the cost savings of supplying multiple units. Allowing package bidding eliminates this risk, potentially leading to more efficient outcomes and lower procurement costs.

On the other hand, allowing package bids can also hurt the performance of a first-price CA. As pointed out by Cantillon and Pesendorfer (2006) and Olivares et al. (2011), bidders can engage in *strategic bundling* in which they submit package discounts even in the absence of cost synergies. One motivation to do so may be to leverage a relative cost advantage in a unit A (for which the bidder is the cost efficient provider) into another unit B (for which the bidder is not the efficient provider), by submitting a discounted package bid for A and B to try to win both units. If the bidder wins the package it will lead to an inefficient allocation where a unit ends up being served by a supplier with a higher cost. In addition, package bidding can also lead to a *free-riding problem* (also known as the *threshold problem*), in which "local" suppliers bidding for small packages free-ride on each other to outbid "global" suppliers submitting bids on larger packages; this free-riding can lead to less competitive bidding and thereby higher payments for the auctioneer (Milgrom, 2000).

Due to the aforementioned trade-off between cost synergies and strategic behavior introduced by package bidding, it is important to analyze the actual performance that CAs have in practice. If cost synergies are strong and the incentives for the types of strategic behavior mentioned above are weak, then a CA should achieve a good performance relative to auction mechanisms that preclude package bidding. On the other

hand, if the strategic motivations alluded above are strong and cost synergies are weak, package bidding may hurt the efficiency and optimality of the auction. Hence, understanding the performance of a CA requires knowledge of cost synergies, as well as the incentives that drive strategic behavior. Unfortunately, existing theory is not conclusive on how large these incentives are in a specific application. Moreover, cost information is typically private and sensitive information of the bidders. Thus motivated, the objective of this paper is to provide an empirical methodology that can be used to evaluate the performance of a first-price sealed-bid CA using bid data. We apply our method to the Chilean school meals combinatorial auction that we describe below.

To measure the performance of a CA, it is essential to identify bidders' supplying costs, which are not directly observable in the bid data. In the context of the application we study, previous work by Olivares et al. (2011) provides evidence of significant package discounts (see Figure 1). Note that even if the bid prices of packages decrease as the package sizes increase, these discounts cannot be directly interpreted as cost synergies. They could also be driven by markup adjustments motivated by the types of strategic behavior alluded above. Although Olivares et al. (2011) provide further suggestive evidence of cost synergies in this application, their approach does not provide direct estimates of the suppliers' costs, which are needed to evaluate the performance of the auction. In addition, their approach does not allow to perform counterfactuals. Other reduced form approaches suffer from the same limitations (see Ausubel et al. (1997), Gandal (1997), and Moreton and Spiller (1998)).

As an alternative to this reduced form approach, we propose a structural estimation approach which directly identifies the bidders' costs using actual bid data. In particular, our structural method disentangles whether the discounts observed in bids are driven by cost synergies or strategic markup adjustments. Our method is based on the seminal work of Guerre et al. (2000) for single-unit auctions that was later extended by Cantillon and Pesendorfer (2006) to a CA setting with a small number of units. The main idea behind this structural approach is to use the first-order conditions from the bidder's profit maximization problem to find the imputed costs that would rationalize the bids observed in the data. Because the bidder's problem involves beliefs about the competitors' bidding behavior, this approach also requires estimating a statistical distribution of the competitors' bids.

In the large-scale CA that we analyze in this paper – where each bidder submits in the order of hundreds or thousands of bids – a direct application of the Cantillon and Pesendorfer (2006) method is not possible due to the large number of decision variables in the bidder's profit maximization problem. We develop a novel approach to overcome this issue, assuming a "simplified" version of the bidder's problem where the markups charged on the package bids are chosen based on a reduced set of package characteristics. With this simplification, the bidder's problem becomes computationally and econometrically tractable so that the structural approach can be effectively applied to large-scale CAs. Recall, however, that the main objective of the structural approach is to identify the cost structure – which is a primitive in the structural model – separately from the markups, which is chosen strategically by the bidders. Therefore, in using our proposed approach it is important to allow for sufficient flexibility in the markup specification so that strategic markup adjustments are not overly restricted in this simplified bidder's problem. We provide detailed guidelines on how to do this in the estimation.

We expect that our approach, based on pricing package characteristics, can be used in several real-

world large-scale auctions. In particular, we effectively apply our method to the Chilean school meals CA (see Epstein et al. (2002) for a detailed description of the auction). This application fits well within the class of large-scale CAs: each auction has about 30 units and firms submit hundreds or even thousands of bids. This CA has a single-round sealed-bid first-price format. The auction is used by the Chilean government to allocate contracts among private catering firms to provide breakfast and lunch for 2.5 million children daily in primary and secondary public schools during the school year. In a developing country where about 14 percent of children under the age of 18 live below the poverty line, many students depend on these free meals as a key source of nutrition. The CA, one of the largest state-run auctions in Chile, was used for the first time in 1999 and has been used every year since its inception awarding more than $3 billion of contracts. Although this application has been praised for bringing transparency and lowering the procurement costs of a high social impact public service, a detailed performance analysis of the CA format has not been conducted.

Our results show that for the Chilean auction, cost synergies are significant, amounting up to 6% of the cost. Roughly 75% of the discounts observed in the bid data arise from cost synergies (the rest is due to strategic markup adjustments). In part due to this large cost synergies, the CA achieves a strikingly high efficiency, with an actual cost allocation only 1% higher than the minimum-cost allocation. The results also show that while economies of scale (mostly generated by volume discounts in input purchases) are larger than economies of density (arising from common logistics infrastructure used to supply nearby units), they are both important in the firms' operational cost synergies. Finally, the estimated markups are on average around 5%, suggesting that the CA induces a reasonable amount of competition among the suppliers. The level of markups coincides with anecdotal evidence provided by the Chilean government. Going back to our initial motivating auction design question, overall, our results suggest that package bidding and running a CA seems appropriate in our setting. Further, the results suggest that the bidding language should allow bidders to express both economies of scale and density.

Once we estimate the cost structure we can also perform other useful counterfactuals. One important consideration the government has when running these auctions, which arises frequently in other settings with synergies, is how to promote diversification and competition among bidders. In the Chilean auction, the government imposes market share restrictions for bidders in the CA to promote long-run competition. The cost estimates provided by the structural estimation can be used to evaluate the efficiency loss due to these constraints. We find that the efficiency loss is very small, around 1%. The main reason for this result is that cost synergies get practically exhausted at the point where the market share constraints become binding.

An important practical motivation to use a first-price rule in applications of CAs is that a Vickrey-Clarke-Groves (VCG) mechanism, that is known to be truthful and efficient, can lead to excessively large payments and other undesirable properties in the presence of synergies (Ausubel and Milgrom, 2006). We conduct a counterfactual experiment to compare the total payments of the first-price sealed-bid CA against a VCG mechanism. Finding the counterfactual total payment of VCG requires identifying the efficient allocation, which can be computed from the cost estimates obtained via the structural estimation. Interestingly, and contrary to the theoretical results mentioned above, we find that in our application the total VCG payment is quite reasonable and very close to the first-price CA payment. We believe this result is driven by the significant amount of competition introduced by the large number of package bids submitted by firms.

Our work is related to other structural estimation papers in auctions (see Athey and Haile (2006), Hen-

dricks and Porter (2007), and Paarsch and Hong (2006) for good surveys). Most notably, Reguant (2011) uses a first-order-conditions approach to structurally estimate a model of the day-ahead wholesale electricity market in Spain. After estimating the cost structure of bidding firms, she performs counterfactuals to determine the welfare effects of "complex bids"; a specific bidding mechanism that allow companies to express cost complementarities of operating across different hours of a day. In addition, Fox and Bajari (2011) use an estimator based on a pairwise stability condition to estimate complementarities in an FCC spectrum auction, which is run in an ascending auction format without package bidding.

There has also been an important literature studying multi-unit auctions of homogeneous goods. We describe a sample of these papers here. Hortaçsu and McAdams (2010) develop a structural estimation approach for the Turkish treasury auction and uses the estimates to compare the performance of a uniform price and a Vickrey auction. Kastl (2011) studies how to use a structural approach when bidders submit discrete bid points rather than continuous downward sloping demand functions with data from the Czech Treasury auction. Finally Chapman et al. (2005) develop a framework to measure best-response violations in multi-unit, sealed-bid, discriminatory-price auctions run by the Bank of Canada to manage excess cash reserves.

Our work is also related to the growing literature in operations management that uses structural estimation. Olivares et al. (2008) develop a structural approach to impute the cost of overage and underage of a newsvendor, which is applied to the reservation of operating room time by an hospital. Allon et al. (2011) conduct a structural estimation to measure the implicit waiting cost of customers in the fast food industry. Similarly, Aksin-Karaesmen et al. (2011) estimate customer waiting costs but develop a dynamic structural model to explain customer abandonments in a bank's call center. Li et al. (2011) also model consumer's forward looking behavior through a dynamic structural model, using data from the airline industry. We add to this stream of research by applying structural estimation in a service procurement setting, an important area in operations and supply chain management where structural estimation methods have not been used.

The rest of the paper is structured as follows. Section 2 develops a structural estimation framework to estimate the primitives of first-price sealed-bid CAs and proposes our structural model for large-scale CAs. Section 3 provides a description of the Chilean auction for school meals and our data set. Section 4 describes the details of our estimation method for the Chilean school meals auction and reports the estimation results. We evaluate the current auction format through efficiency analysis and perform counterfactual analysis in Section 5. Section 6 describes our main conclusions.

2 Structural Estimation Framework

This section develops a structural estimation framework to estimate the primitives of large-scale first-price single-round sealed-bid CAs. First, in Section 2.1 we describe a general structural estimation framework for CAs. A similar approach has been successfully implemented by Cantillon and Pesendorfer (2006) (hereon CP) to estimate small London bus route CAs. This approach is itself inspired by the pioneering work by Guerre et al. (2000) for single-unit auctions. However, there are limitations of using this approach when the number of units and possible packages grows, which is common for many CAs in practice. Section 2.2 describes our contribution to address these limitations, which basically relies on assuming a simplified bidding

strategy that reduces the complexity of the bidder's problem, making the structural estimation feasible for large-scale CAs. Section 2.3 provides guidelines on how to implement our structural estimation approach and Section 2.4 discusses important identification issues.

2.1 A Structural Estimation Approach to First-Price Sealed-Bid Combinatorial Auctions

We begin by describing a structural estimation approach to first-price single-round sealed-bid CAs. The approach is similar to the one introduced by CP.

First, we describe the basic setting of a CA. Let U denote the set of N units to be procured by an auctioneer. There is a set F of supplier firms, referred to as bidders and indexed by f. A package or combination, indexed by a, is a non-empty subset of units in U. We let $\mathcal{A}$ denote the set of all possible packages and $A = |\mathcal{A}| = 2^N - 1$ be the total number of possible packages. Let b_{af} denote the bid price asked by bidder f to supply package a, and $b_f = \{b_{af}\}_{a \in \mathcal{A}}$ the *bid vector* containing all bids from that bidder.

The following assumption describes the auction format.

Assumption 1 (Auction Format). *The auction has a first-price single-round sealed-bid format, so that bidders submit their bids simultaneously and winning bidders are paid their submitted bid prices for the packages awarded to them. The auction mechanism determines the winning bids by solving the following mathematical integer program:*

$$\begin{aligned} \min \quad & \sum_{a \in \mathcal{A}, f \in F} b_{af} x_{af} \\ \textit{s.t.} \quad & x \in X, \quad x_{af} = \{0,1\}, \, \forall a \in \mathcal{A}, f \in F, \end{aligned} \tag{1}$$

where x_{af} is a binary decision variable that is equal to one if package a is assigned to bidder f, and $x = \{x_{af}\}_{a \in \mathcal{A}, f \in F}$. We denote by X the set of feasible allocations; the set imposes that each unit is allocated to one bidder, that each bidder can win at most one package, and potentially some additional allocative constraints.

The winner determination problem minimizes the total procurement costs of the auctioneer, given the submitted bids. We note that the additional constraints in the set of feasible allocations could impose, for example, market share constraints that limit the maximum package size that a single bidder can be awarded, which may be used to keep a diversified supplier base. In Section 3.2 and the online appendix we provide more details on the mathematical integer program that solves the winner determination problem in the context of our specific empirical application.

The structural estimation approach requires assumptions on the bidders' information structure and bidding behavior in order to identify costs. We make the following assumption.

Assumption 2 (Bidders' Costs). *Bidders have independent private costs. In particular, given an auction, each bidder gets an independent random draw of a cost vector $c_f = \{c_{af}\}_{a \in \mathcal{A}}$, where c_{af} is the cost of supplying package a for bidder f.*

Before submitting its bid, each bidder observes its own vector of costs, but does not observe the costs' realizations of its competitors. Moreover, because costs are private, a bidder's costs only depend on its own private signal and it is not a function of the costs' realizations of other bidders. We make the following assumption on bidders' strategies.

Assumption 3 (**Strategies**). *Bidders are risk-neutral and play pure bidding strategies. A bidder's strategy is a function* $b_f : \Re_+^A \mapsto \Re_+^A$ *that depends on its own costs* c_f*. Bidders place bids on all possible combinations of units.*

In our sealed-bid format, bidders submit their bids in a game of asymmetric information without directly observing the bids nor the cost realizations of their competitors. Therefore, bidders face uncertainty on whether they will win any given package. For each bidder, we capture this uncertainty with the vector $G_f(b_f) = \{G_{af}(b_f)\}_{a\in\mathcal{A}}$, where $G_{af}(b_f)$ is the probability bidder f wins package a with bid vector b_f. Using vector notation, we can then write a bidder's expected profit maximization problem as:

$$\max_{b\in\Re^A} \quad (b-c)^T G(b), \tag{2}$$

where v^T denotes the transpose of a vector v. Note that each bidder has its own optimization problem with its own cost and winning probability vectors. Whenever the context is clear, we omit the subscript f to simplify the notation.

To formulate the optimization problem above, a bidder needs to form expectations about the bidding behavior of its competitors, so that it can evaluate the vector of winning probabilities $G(b)$, for a given value of b. Note that if bidder f anticipates that bidder f' uses a bidding strategy $b_{f'}(\cdot)$, the bids of bidder f' are random from bidder f's perspective; they correspond to the composition $b_{f'}(c_{f'})$, where $c_{f'}$ is the random cost vector for bidder f'. Assumption 4, described next, formalizes this. Assumptions 1, 2, 3, and 4 are kept throughout the paper.

Assumption 4 (**Bid Distributions**). *a) Consider a given auction and any bidder* f*. From its perspective, assume that competitors' bid vectors* $b_{f'}$ *are drawn independently from distributions* $H(\cdot|Z_{f'})$*, where* $Z_{f'}$ *is a vector of observable characteristics of bidder* f'*. These distributions are common knowledge among bidders and induce the correct vector of winning probabilities* $G_f(b_f)$*, for all* b_f*, given the competitors' strategies and the cost vectors' probability distributions.*
b) In addition, assume that for all bidders $f \in F$*,* $H(\cdot|Z_f)$ *has a continuous density everywhere.*

Note that the independence part of the assumption is consistent with Assumptions 2 and 3. Also note that Assumption 4 captures all the relevant uncertainty faced by the bidder when solving (2). In particular, for a given bid vector b_f submitted by bidder f, the competitors' bid distributions $\{H(b_{f'}|Z_{f'})\}_{f'\neq f}$ and the allocation rules given by the winner determination problem uniquely determine winning probabilities of each bid $G_{af}(b_f)$.

Previous work, like CP and Guerre et al. (2000), assume that the primitives of the model such as the number of bidders, the probability distribution of costs, and the utility functions are common knowledge and that bidders play a Bayes Nash equilibrium (BNE) of the game induced by the auction. In many settings,

such as the first-price single-item auction studied in Guerre et al. (2000) this is well justified; under mild conditions a unique symmetric BNE always exist. However, there are no theoretical results that guarantee existence, uniqueness, nor characterization of equilibrium for a CA. Hence, we make slightly weaker assumptions, but that still allow us to develop a structural estimation approach.

More specifically, note that assuming BNE play imposes two conditions: (i) bidders correctly anticipate the strategy of their competitors, and therefore correctly estimate the vector of winning probabilities given their own bids; and (ii) for each bidder, given its costs and the winning probabilities function, the bidder selects a bid vector that maximizes its expected profit. While Assumption 4.(a) is weaker than condition (i), it imposes the same restriction over bidders' beliefs that we use in our structural estimation approach: bidders in the auction can correctly anticipate their winning probabilities. We also make a weaker assumption relative to the aforementioned condition (ii) imposed by BNE: we will only assume that bidders select a bid vector that satisfies the necessary first-order conditions of the expected profit maximization problem, and these are not sufficient for optimality in a CA. We come back to this point in the sequel.

Assumption 4.(b) guarantees the differentiability of the winning probability vector $G(\cdot)$ that is needed to use the first-order conditions for estimation, as we formalize in the next lemma. Note that this assumption is over the bids' distribution, that is endogenous in the auction game. Although we would prefer to make assumptions over model primitives that imply the assumptions on behavior, the lack of theoretical results regarding the existence and characterization of equilibrium in CAs does not allow us to follow this approach.

Proposition 1. *Consider a given auction. For every bidder, the winning probability vector $G(b)$ is continuous and differentiable, for all b.*

The proof of this proposition as well as all other proofs are provided in Online Appendix B. For a given bidder, the necessary first-order conditions of the optimization problem (2) are given by the following vector equation:

$$c = b + \left\{ [\mathcal{D}_b G(b)]^T \right\}^{-1} G(b), \tag{3}$$

where $\mathcal{D}_b$ refers to the Jacobian matrix operator with respect to the variable vector b so that the ij^{th} element is $[\mathcal{D}_b G(b)]_{ij} = \frac{\partial}{\partial b_j} G_i(b)$. The Jacobian is a square matrix which can have non-zero off-diagonal elements because packages of the same bidder compete against each other. Note that for a given auction there is one first-order condition vector equation per bidder; these equations are the basis to identify the cost vectors of each bidder, as we now explain.

Notice that, for a given bidder, the right-hand side of equation (3) only depends on the observed bid vector b, the winning probabilities $G(b)$ and its derivatives. By Assumption 4, the vector of winning probabilities $G(b)$ must be consistent with the actual auction play observed in the data, and therefore can be potentially estimated using bidding data from all bidders. For example, Guerre et al. (2000) estimate the distribution $G(\cdot)$ (and its derivative), which in a single-unit procurement auction corresponds to the tail distribution of the minimum bid, using a non-parametric approach.

In a CA setting, $G(\cdot)$ is a vector of probabilities, which complicates its estimation. A possible approach to estimate $G(\cdot)$ is to parametrically estimate the bid distribution of competing bidders ($H(\cdot|Z_f)$ in Assumption 4) using bidding data. Note that this is a highly dimensional distribution so the parametric as-

sumptions will be important to make the estimation tractable. In Section 4.1 we provide more details about a parsimonious, yet flexible parametric description of this distribution in the context of our application.

Note also that previous structural approaches usually use a cross section of auctions for the estimation of $G(\cdot)$, assuming that in all auctions the same equilibrium is being played. In our approach we conduct an auction-by-auction estimation which does not require the latter assumption. Also, our estimates are less susceptible to unobserved heterogeneity across auctions, that is, to the effect of characteristics observed by the bidders but not by the econometrician that we do not control for in $\{Z_f\}_{f \in F}$. As we discuss later, for a given auction, we exploit the large number of units and packages in our application to parametrically estimate the distribution of competitors' bids.

Using the distribution of competitors' bids, one can use simulation to estimate the winning probabilities by sampling competitors bids from these distributions and solving the winner determination problem repeatedly. Derivatives could then be computed using a finite difference method. In fact, CP uses this method and replaces the estimates of $G(b)$ and $\mathcal{D}_b G(b)$ together with the observed bid vector b in equation (3) to obtain an estimate of c for each firm. They were able to effectively use this approach in auctions of at most 3 units.

Even if one was able to parametrically estimate the distribution of competitors' bids, there is an important limitation of the previous approach in larger-scale CAs: the dimensionality of the optimization problem (2) increases exponentially with the number of units. For example, in our application there are millions of possible packages and bidders submit in the order of hundreds or thousands of bids. Let us revise the first-order condition (3) in this context. First, for a given bidder, we need to estimate hundreds or thousands of winning probabilities. As the number of bids submitted by a bidder increases, the winning probability of each bid is likely to become very small and the simulation error of these low-probability events becomes large. Moreover, equation (3) requires taking derivatives over a large number of variables; simulation error for these quantities may be even larger. Hence, computation of $G(b)$ and $\mathcal{D}_b G(b)$ becomes quickly intractable as the number of units auctioned increases.

The difficulties in estimating $G(b)$ make it also unreasonable to assume that bidders would be able to solve (2) optimally. One approach to simplify the bidders' problem to make it more amenable for analysis is to reduce the set of decision variables. The next section describes a structural model which incorporates this simplification.

2.2 The Characteristic-Based Markup Approach for CAs

Our model is based on the bidder's problem (2), which we refer to as the *full-dimension* problem in the sense that the bidder chooses every bid price. As mentioned above, the main complication of using this model in a large-scale CA is that the dimension is too large. In what follows, we present an approach to reduce the dimensionality of the problem. In particular, we develop a structural estimation approach that imposes additional assumptions on the bidders' bidding behavior that have behavioral appeal and make the estimation approach econometrically and computationally feasible in large-scale CAs.

Notice that the first-order condition (3) can be re-written as $b = c + \left(- \left\{ [\mathcal{D}_b G(b)]^T \right\}^{-1} G(b) \right)$, so that the bid is a cost plus a markup. Hence, we can view the full-dimension problem as choosing a markup for each package. We propose instead that this markup is specified by a reduced set of package characteristics.

Specifically, let w_a be a row vector of characteristics describing package a, with dimension $\dim(w_a) = d$ much smaller than A. The markup for package a is given by the linear function $w_a\theta$, where θ is a (column) vector of dimension d specifying the markup associated with each package characteristic. Instead of choosing the markup for each package, the bidder now chooses θ – the set of markups associated with each of these reduced set of package characteristics. Let $W \in \Re^{A\times d}$ be a matrix containing the characteristics of all packages, so that the a^{th} row of W is w_a. The following assumption, kept throughout the paper, formalizes this simplification to the bidders' bidding behavior.

Assumption 5 (**Characteristic-Based Markups**). *Consider a given bidder in a particular auction. Its bid vector is determined by $b = c + W\theta$, where W is a fixed $(A \times d)$- dimensional matrix of package characteristics and θ is a d-dimensional decision vector chosen by the bidder.*

Note that different bidders can adopt different W matrices. Under this assumption, the bidder's optimization problem becomes:

$$\max_{\theta\in\Re^d} \quad (W\theta)^T G(W\theta + c), \tag{4}$$

whose first-order conditions yield:

$$[\mathcal{D}_\theta W^T G(W\theta + c)]^T\theta = -W^T G(W\theta + c). \tag{5}$$

Here again the ij^{th} element of the Jacobian matrix above is $[\mathcal{D}_\theta W^T G(W\theta+c)]_{ij} = \frac{\partial}{\partial\theta_j}[W^T G(W\theta+c)]_i = \frac{\partial}{\partial\theta_j}W_i^T G(W\theta + c)$, where W_i is the i^{th} column of matrix W. Re-arranging and replacing terms, we can solve for the decision vector θ as follows:

$$\theta = -\left\{[\mathcal{D}_\theta W^T G(b)]^T\right\}^{-1} W^T G(b). \tag{6}$$

As in Guerre et al. (2000) and Cantillon and Pesendorfer (2006), this first-order condition equation constitutes the basis of identification in our structural model. Again, note that in each auction there is one first-order condition vector equation per bidder and, for each bidder, under Assumption 5, the cost is given by $c = b - W\theta$. Hence, costs are uniquely determined by θ, and, moreover, if the matrix $\mathcal{D}_\theta W^T G(b)$ is invertible, equation (6) uniquely identifies the markup vector θ. Hence, equation (6) provides an alternative to (3) to estimate costs. In Section 2.4 we study conditions for the invertibility of this matrix and for identification. We formalize this discussion with the following assumption that is kept throughout the paper.

Assumption 6 (**First-Order Conditions**). *The observed bid vector of a given bidder in the auction satisfies the necessary first-order conditions of the characteristic-based markup model given by* (5)*.*

Note that the bidder's optimization problem in a CA is not necessarily concave. Hence, the first-order conditions (5) are not sufficient for optimality for the reduced optimization problem (4). Despite that, it is in principle possible to test computationally whether the observed bid vector that satisfies (5) is locally or globally optimal for optimization problem (4). We provide more details in the context of our application.

Similarly to equation (3), the right-hand side of equation (6) can be estimated purely from observed bidding data when it is evaluated at the observed bid vector b. Basically, the winning probability vector $G(b)$

and its Jacobian matrix $\mathcal{D}_b G(b)$ in the full-dimension model are replaced by the vector $W^T G(b)$ and its Jacobian matrix $\mathcal{D}_\theta W^T G(b)$ which is now with respect to the markup variable vector θ. Equation (6) provides substantial computational advantages to the estimation process compared to equation (3). First, derivatives are now taken with respect to $d \ll A$ variables, effectively reducing the dimension of the problem. Second, instead of estimating the winning probability $G(b)$ and its Jacobian matrix, it is enough to estimate $W^T G(b)$ and its Jacobian matrix $\mathcal{D}_\theta W^T G(b)$. Later we show examples that illustrate how $W^T G(b)$ aggregates probabilities over many packages. Hence, there are much fewer probabilities to estimate, and each one has a larger value so they are easier to estimate than winning probabilities of individual packages. This makes the estimation tractable.

One apparent limitation of Assumption 5 is that the markup is additive as oppose to multiplicative to costs, which may be more appropriate. A multiplicative markup, however, would lead to different first-order conditions from which it is mathematically intractable to identify bidders' costs using bid data. A relatively simple way to make the additive assumption less restrictive is to include package characteristics in W which are related to costs, so that the markup can be scaled based on these cost-characteristics. This approach is effective when the cost heterogeneity across packages can be captured by a reduced set of known variables.

The characteristic-based markup model is very general and flexible in the specification of markup structures. For example, if we specify the package-characteristic matrix W as the identity matrix, each package has its own markup and we are back to the full-dimension problem (2). On the opposite extreme, we could choose $d = 1$ so that all packages share the same markup, reducing the problem to a single decision variable. Between these two extremes there are many possible specifications for W. Note that data can provide guidance on what is a reasonable specification. For example, the solution of the full-dimensional first-order conditions (3) would provide information on what packages have similar markups and on what package characteristics affect markups the most. However, as we mentioned, solving (3) is intractable. In the next section we describe an alternative approach that uses the data in a tractable way to determine a reasonable specification for W that balances computational efficiency versus flexibility in the markup structure.

2.3 Specifying the Package-Characteristic Matrix W

Recall that our main objective is to identify how much of the discounts observed in the bids are due to cost synergies as opposed to strategic markup adjustments when bidding for larger packages. In order for the characteristic-based markup approach to capture this type of strategic markup adjustments, it is important to allow the markup to vary in the size of a package. For this reason, we introduce a package-characteristic matrix W that includes variables related to the size of each package. Doing so helps separating what portion of the volume discounts observed in the bid data arises from markup adjustments vis-à-vis cost synergies.

We describe how to incorporate the size of a package as a characteristic in W. Let $\{\mathcal{A}_s\}_{s=1}^S$ form a partition of the set of possible combinations $\mathcal{A}$ that groups combinations in terms of some size measure. For example, this partition could be specified in terms of the number of units in a package. Hence, θ_s represents the markup charged for any package of size s, and so the bidder needs to choose S different markups, one for each possible size. The package-characteristic matrix $W \in \Re^{A \times S}$ can be specified by the indicator variables $W_{as} = \mathbf{1}[\text{package } a \text{has size } s]$. With this specification, the term $W^T G(b)$ in equation (6) has the

following form:

$$W^T G(b) = \begin{bmatrix} W_1^T G(b) \\ W_2^T G(b) \\ \vdots \\ W_S^T G(b) \end{bmatrix} = \begin{bmatrix} \text{Probability of winning any package of size 1} \\ \text{Probability of winning any package of size 2} \\ \vdots \\ \text{Probability of winning any package of size } S \end{bmatrix}$$

The previous *size-based markup* model significantly reduces the dimensionality of the problem making the estimation feasible. In particular, while the winning probability of any given package a is typically small and hard to estimate via simulation, the winning probability of a *group* of packages of the same size is a sum of these individual probabilities over a potentially large set of packages, and may be much larger. This makes the computation of the right-hand side of the first-order condition (6) tractable.

On the other hand, the size-based markup model may be too restrictive. It may be the case that two packages of the same size would have significantly different markups in the full-dimension model and should not be grouped together in the size-based markup model. We now study how these restrictions on setting markups may affect the estimation. To do this, we provide an analytical comparison of the markups estimated by the full-dimension model with those estimated via the characteristic-based markup approach. We use these results to develop a heuristic which uses the data to construct a reasonable package-characteristic matrix W that refines the size-based markup model, balancing flexibility in the markup structure with computational efficiency in the estimation.

Consider a situation in which the markups of K packages are aggregated into a single common markup. More formally, in the full-dimension model we have $b_a = c_a + \theta_a,\ a = 1, ..., K$ and in this specific characteristic-based model we have $b_a = c_a + \theta_u,\ a = 1, ..., K$, where θ_u is the common markup. We consider the perspective of a specific bidder and show the following proposition.

Proposition 2. *Consider a given bidder placing a bid vector b in a CA with K packages. Suppose the bids on the K packages have positive winning probabilities and let $\theta_a,\ a = 1, ..., K$ be the solution of the first-order condition of the full-dimensional model,* (3)*. Let θ_u be the common markup for all K packages that solves the first-order conditions of the characteristic-based model,* (6)*. Then,*

$$\begin{aligned} \theta_u &= \frac{1}{\sum_{a=1}^K \alpha_a} \sum_{a=1}^K \alpha_a \theta_a, \text{ where} \\ \alpha_a &= \frac{\partial G_a(b)}{\partial \theta_u} = \sum_{s=1}^K \frac{\partial G_a(b)}{\partial \theta_s}. \end{aligned}$$

Moreover, $\alpha_a \leq 0,\ \forall a$, with at least one $\alpha_a < 0$.

Note that because α_a's are negative, the common markup θ_u is a weighted average of the individual markups θ_a's. From this we learn that, if we (the researchers) use the uniform markup model while the bidder actually solves the full-dimension model, the identified markup will be a weighted average of the individual markups. Hence, if groups of packages have similar markups in the full-dimension model, using these

groups as characteristics in the W matrix would make the reduced markup model a good approximation.[1]

Proposition 2 shows the impact on the estimated markup of those packages that are grouped. It is also important to analyze how the grouping affects other packages not contained in the group. We begin considering a two unit case where we impose a common markup, θ_u, for the stand-alone bids containing a single unit. The idea is to study how this grouping affects the estimated markups of the package of two units. The following proposition summarizes this result.

Proposition 3. *Consider a CA with 2 units and a given bidder. Suppose that all the bids have positive winning probabilities and that* $(\theta_1, \theta_2, \theta_{12})$ *solves the first-order conditions of the full-dimensional model,* (3)*. Suppose* (θ_u, θ_v) *is the solution of the first-order conditions of the characteristic-based model,* (6)*, where* θ_u *is the common markup for single unit bids and* θ_v *is the markup for the package bid. Then,*

$$\theta_u = \lambda\theta_1 + (1-\lambda)\theta_2,$$
$$\theta_v = \theta_{12} + \gamma(\theta_1 - \theta_2),$$

where

$$\lambda = \left\{ \frac{\partial G_{12}}{\partial \theta_{12}} \left(\frac{\partial G_1}{\partial \theta_1} + \frac{\partial G_1}{\partial \theta_2} \right) - \frac{\partial G_1}{\partial \theta_{12}} \left(\frac{\partial G_{12}}{\partial \theta_1} + \frac{\partial G_{12}}{\partial \theta_2} \right) \right\} \Big/ det,$$
$$\gamma = \left\{ \left(\frac{\partial G_2}{\partial \theta_1} + \frac{\partial G_2}{\partial \theta_2} \right) \frac{\partial G_1}{\partial \theta_{12}} - \left(\frac{\partial G_1}{\partial \theta_1} + \frac{\partial G_1}{\partial \theta_2} \right) \frac{\partial G_2}{\partial \theta_{12}} \right\} \Big/ det,$$
$$det = \left(\frac{\partial G_1}{\partial \theta_1} + \frac{\partial G_2}{\partial \theta_1} + \frac{\partial G_1}{\partial \theta_2} + \frac{\partial G_2}{\partial \theta_2} \right) \frac{\partial G_{12}}{\partial \theta_{12}} - \left(\frac{\partial G_{12}}{\partial \theta_1} + \frac{\partial G_{12}}{\partial \theta_2} \right) \left(\frac{\partial G_1}{\partial \theta_{12}} + \frac{\partial G_2}{\partial \theta_{12}} \right).$$

From the above proposition we again observe that θ_u is a weighted average of the individual markups. Moreover, we also observe that grouping the unit markups affects the estimated markup of the package. However, as seen in the above equation, the impact of grouping on the bundle markup, the one that is not grouped, depends on the coefficient γ and the difference of the grouped markups. If the unit markups are very close to each other, the effect of grouping will be negligible. Quantifying the value of γ analytically is a challenging task due to the limited knowledge on the winning probabilities. Nevertheless, computationally we have observed that in practice γ has low values. Therefore, we expect that the effect of grouping markups on out-of-group markups will be small. In fact, our numerical experiments have shown that in our application, grouping a set of packages so that they share a common markup merely affects the markups of other packages not in this group.

Our previous discussion and Propositions 2 and 3 provide some guidance on how to construct a reasonable package-characteristic matrix. First, we only consider packages of the same size to form a group with a common markup. Second, we try to group markups so that the packages in the same group would have similar markups in the full-dimension model. Packages that would have significantly different markups compared to the rest of the group in the full-dimension model should be separated and have their own markup. Ideally, one could use the first-order conditions of the full-dimension model (3) to identify such

[1] We note that to obtain the result that $\alpha_a \leq 0, \forall a$, it is important to assume that $\{\theta_a\}_{a\in\{1,\ldots,K\}}$ and θ_u are absolute markups, i.e., markup values for corresponding units or packages. Hence, elements in the package-characteristic matrix W are either zero or one.

packages, but this is intractable. Instead we use the packages' winning probabilities as proxies for markup values. In fact, while this proxy is not perfect, numerical experiments suggest that packages with larger winning probabilities are likely to have larger markups in the full-dimension model. Finally, we require a minimum threshold on the aggregated winning probabilities of each package group. Recall that to estimate the markup using equation (6), we need to estimate the aggregated winning probabilities and its derivatives numerically. The minimum probability threshold is applied to ensure accurate estimation of those terms.

Based on these ideas we develop a heuristic to build the package-characteristic matrix for a given firm. The method roughly consists of three steps:

1. First, we run a simulation to estimate winning probabilities of each package; this simulation is quicker to run than solving for the first-order conditions. We identify packages that have very high winning probabilities relative to the rest. These packages are likely to have larger markups in the full-dimension model, so each of them is associated with its own markup variable. For the rest of the packages, we form several groups of packages so that packages with similar winning probabilities are grouped together. In addition, recall that we only group together packages of the same size. In short, in this step, we try to have as many markup groups as possible to the extent that computational tractability is maintained.

2. In the second step, given the candidate matrix W (or equivalently, the package groups) constructed from the first step, we obtain rough estimates for the markups using equation (6). Given those estimates, we combine some of the groups together if they have similar markup levels. As suggested in the propositions of the previous subsection, this will increase computational tractability without sacrificing the flexibility of the markup structure too much.

3. This gives us the final choice of the package-characteristic matrix W, with which we obtain precise and final estimates of the markup vector θ through equation (6).

We use this heuristic to build the W matrix for each bidder in our estimation method.

2.4 Identification

A condition to uniquely identify the markup vector θ, and hence the costs, is that the matrix $\mathcal{D}_\theta W^T G(b)$ is invertible in equation (6). We finish this section discussing issues related to identification which are important for the specification of W.

In empirical settings, including the one analyzed in this paper, bidders may not submit bids on all packages.[2] These unobserved package bids can still be incorporated into our framework (which assumes bids for all packages) by treating them as observed bids with very high prices that have no chances of winning. We refer to the bids which never win as *irrelevant bids*. In addition, some bids that are actually submitted may also be irrelevant, in the sense that they have zero probabilities of winning. For example,

[2] In fact, in our empirical application, firms do not place bids on all possible combinations because of two reasons: (1) firms have limits on the maximum number of units that can be included in a package (these limits depend on the firm's financial capacity); and (2) the number of possible combinations is too large.

this could arise as a strategic decision in order not to win a specific package when the auction rules require submission of bid prices on all packages. In what follows, we show how irrelevant bids can limit the identification of costs and how they can be handled in an actual application. We also provide necessary and sufficient conditions so that (6) identifies markups, and thereby costs.

Recall that each column of package characteristics in W is associated with a markup variable in the bidder's decision θ. We say that a package a is associated with the markup variable θ_i if $W_{ai} \neq 0$, that is, the bid price of a depends on the value of θ_i. The following lemma is useful to characterize the conditions needed for identification:

Lemma 1. *Consider a given bidder and auction. For any package $a \in \mathcal{A}$, $G_a(b) = 0$ implies $\frac{\partial}{\partial \theta_i} G_a(W\theta + c) = 0$, for all $i = 1, ..., d$.*

The lemma implies that if all the bids associated with a markup variable θ_i are irrelevant, then the i^{th} row of $\mathcal{D}_\theta W^T G(b)$ matrix will be all zero, and the matrix will not be invertible. Because equation (6) requires invertibility of the Jacobian, the markup vector of that bidder is not identified. This problem, however, can be resolved by eliminating irrelevant bids from the model. By doing so, we can still identify markup variables as long as they have at least some relevant bids that are associated with it. We examine this issue in more detail in what follows.

Consider a given firm. Without loss of generality, we assume packages are ordered such that all the relevant bid packages (superscripted by R) are followed by the group of irrelevant bid packages (superscripted by I), so that:

$$W = \begin{bmatrix} W^R \\ \cdots \\ W^I \end{bmatrix}, \quad c = \begin{bmatrix} c^R \\ \cdots \\ c^I \end{bmatrix}, \quad b = \begin{bmatrix} b^R \\ \cdots \\ b^I \end{bmatrix}, \text{ and } \quad G(b) = \begin{bmatrix} G^R(b) \\ \cdots \\ G^I(b) \end{bmatrix}$$

Replacing in equation (6) we obtain:

$$\begin{aligned} \theta &= -\left\{ \left[\mathcal{D}_\theta \left((W^R)^T G^R(b) + (W^I)^T G^I(b) \right) \right]^T \right\}^{-1} \left((W^R)^T G^R(b) + (W^I)^T G^I(b) \right) \\ &= -\left\{ \left[\mathcal{D}_\theta (W^R)^T G^R(b) \right]^T \right\}^{-1} (W^R)^T G^R(b) \\ &= -\left\{ \left[\mathcal{D}_\theta (W^R)^T G^R(b^R) \right]^T \right\}^{-1} (W^R)^T G^R(b^R), \end{aligned} \tag{7}$$

where the second to last equation follows from $G^I(b) = 0$ and Lemma 1. In the last equation, it is implicitly assumed that the bidder only submit bids for relevant bids. Because irrelevant bids never win and by Lemma 1 small changes in the markup vector will not turn them into relevant bids, it is the same as if the bidder would not have submitted them (recall that non submitted bids are also irrelevant). Therefore, the right-hand side of equations (6) and (7) are equivalent. Consequently, the elimination of irrelevant bids will not affect the identification of the markup vector θ as long as the Jacobian in equation (7) is invertible.

Like the size-based markup model and its refinement described in the previous section, when the markup of each package is determined by one and only one markup parameter, we call it a *group markup model*. The following theorem provides necessary and sufficient conditions to ensure identification of the markup vector θ for the class of group markup models.

Theorem 1. *Consider a given bidder and auction. Assume that the package-characteristic matrix W is the specification of a group markup model. If the Jacobian matrix $\mathcal{D}_\theta W^T G(b)$ evaluated at the observed bid vector b is invertible, then every markup variable has at least one relevant bid associated with it. The latter condition becomes sufficient for the invertibility of the Jacobian matrix if the following additional conditions hold: (i) the observed bid vector satisfies equation* (5)*; (ii) the observed bid vector is such that $b - c \geq 0$; and iii) all elements of W are nonnegative ($W \geq 0$). In this case, the markup vector θ is uniquely identified by equation* (6)*.*[3]

Note that the assumption $b - c \geq 0$ is a mild rationality assumption on bidders' behavior that guarantees bidders make positive profits on each package conditional on winning that package. Also note that under the previous assumption, assuming $W \geq 0$ is essentially done without loss of generality, in the sense that for a W matrix with negative entries, one can show there is another W matrix with non-negative entries that produces the same markup estimates. A practical implication of the theorem is that when implementing the heuristic described in Section 2.3 one needs to make sure that each group of packages must include at least one relevant bid. After we imposed this, we were always able to invert the Jacobian matrix computationally.

It is important to note that, for a given bidder, our approach only allows us to identify the cost structure of packages associated with relevant bids, that is, $c^R = b^R - W^R\theta$. In fact, irrelevant bids provide no information to the first-order conditions. Although it is not possible to point identify the costs of irrelevant bids, it can be shown that bounds on the costs of such "irrelevant" bid packages can be obtained. CP show that by finding the threshold bid price over which the bid becomes relevant, we can identify a lower bound on the cost of the specific irrelevant bid package. However, in a large-scale CA, this is not viable because of the computational burden. Instead, we infer the costs of those irrelevant bids using extrapolation. We will come back to this point in Section 4 in the context of our application.

Finally, an important assumption needed for our approach is that bidders can win at most one package. This is a frequent requirement in many real-world CAs. Without this requirement, it may not be possible to point identify costs, as we illustrate with the following example. Consider a CA with 2 units and suppose a bidder only submits bids for the individual units. Suppose the bidder has a positive chance of winning both individual bids simultaneously, which is equivalent to winning the two-unit package. Then, we have three unknowns to estimate (the cost for each individual unit and the cost for the package), but only two equations (the two first-order conditions with respect to the individual bid prices).

3 Application: The Chilean Auction for School Meals

The application we study in this paper is the Chilean auction for school meals. In this section, we provide a detailed description of the auction as well as of the data available.

[3]The following two conditions are necessary for identification for a general package characteristic specification: (i) every markup variable has at least one relevant bid associated with it; and (ii) the package-characteristic matrix W^R has full column rank. We have not explored general sufficient conditions for identification beyond the group markup model.

3.1 Brief History

Junta Nacional de Auxilio Escolar y Becas (JUNAEB) is a government agency in Chile that provides breakfast and lunch for 2.5 million children daily in primary and secondary public schools during the school year. This is one of the largest and most important social programs run by the Chilean government. In fact, in a developing country where about 14 percent of children under the age of 18 live below the poverty line, many students depend on these free meals as a key source of nutrition.

Since 1999 JUNAEB assigns its school meal service contracts through a single-round, sealed-bid, first-price CA, that was fully implemented for the first time that year. The CA has been used every year since its inception awarding more than US$3 billion of contracts (US$ 577 million were awarded in 2008), being one of the largest state auctions in Chile.

For the purpose of the auction, Chile is divided into approximately 100 school districts or territorial units (TUs) in 13 geographic regions. JUNAEB holds auctions in one-third of the country every year, for around 30 - 35 TUs each time, awarding three-year contracts. Approximately 20 firms participate in each auction. Firms can submit bids on various groupings of TUs defining the combinatorial character of this auction. This mechanism is motivated by the belief that firms are subject to cost synergies that arise from operational advantages when serving multiple TUs. More specifically, suppliers may face economies of scale (generated by volume discounts in their input purchases) and economies of density (arising from common logistics infrastructure used to supply nearby units).

3.2 Auction Process

The auction process begins when JUNAEB invites and registers potential vendors. The agency then evaluates the companies from a managerial, technical and financial point of view, and eliminates those that do not meet minimum reliability standards. Qualifying vendors are classified according to two characteristics: their financial capacity (based on data from the firms' balance sheets), and their managerial competence. Usually, firms below a minimum level of managerial competence are not allowed to participate in the auction. Meal plans are standardized and service quality requirements are presented in detail. With that, firms compete on price basis. Potential vendors submit their bids simultaneously and in a single-round through an online system. Upon winning a contract, the firm receives its bid as a payment and it is responsible for managing the entire supply chain associated with all meal services in the corresponding TUs. This includes from sourcing food inputs going all the way to cooking and serving the meals in the schools.

Bidding language. A bid can cover any combination from one to eight TUs and specifies the price for which the firm would serve all meals included in the TUs in the combination. Vendors can submit many bids and each package bid is either fully accepted or rejected (i.e. the mechanism does not allocate a fraction of a bid); most firms submit hundreds or even thousands of bids.

Winner determination. The allocation is chosen by selecting the combination of bids that supply all of the TUs at a minimum cost. The problem is formulated as an integer program (IP) that incorporates other considerations and side constraints. There are four types of constraints implemented in the auction and the

details of those constraints are as follows. The mathematical formulation of the IP is provided in Online Appendix A.

1. **Cover all TUs:** the final allocation should cover all the TUs auctioned.

2. **Maximum Number of TUs:** There is a maximum number of TUs that each firm can be allocated in any given auction. This maximum is based on the financial evaluation conducted by JUNAEB every year and therefore can be different across firms and auctions, ranging from 1 to 8 TUs.

3. **Global Market Share Constraints:** To avoid excessive concentration and encourage diversification, at any point in time, the total standing contracts of any firm cannot exceed 16% of the total number of meals included in all TUs in the entire country. Hence, depending on the volume of standing contracts, the maximum volume can be also different across firms and auctions.

4. **Local Constraints:** To facilitate supervision and control of the firms, there are constraints on the maximum number of firms serving in each geographical region. On the other hand, to actively respond to contingencies such as bankruptcies, there are also constraints on the minimum number of firms serving in each geographical region. Geographical regions in low population areas contain less than five TUs while regions with higher population typically contain between 10 and 20 TUs.

5. **Global Competition Constraint:** For similar reasons as the global market share constraints, there is a constraint in the minimum number of firms winning contracts in each auction (this number can vary across auctions, but is around 10).

Finally, we note that in the structural method described in Section 2.1 we assumed that the auction allocates at most one package per firm. While this restriction is not imposed in our empirical application, except for isolated exceptions, firms actually win at most one package in practice. Hence, this is a reasonable assumption even in our application, so we also impose the one-package per-firm constraint in the winner determination problem.

3.3 Description of the Data

We collected data on all auctions between 1999 and 2005. Our data set contains all bids placed by all firms in each auction and the identity of winning firms in each auction. In addition, we also have detailed information on the auction parameters and characteristics of all participating firms. A more detailed description of the data can be found in Olivares et al. (2011).

We have detailed information on the parameters of each auction, including the TUs auctioned. For each TU, we know its annual demand in terms of number of meals to be served and the geographic location of its schools. TUs are heterogeneous in terms of size and the density of the school population, which are key factors affecting firms' supplying costs. In addition, we have all the parameters used in the constraints in the integer program associated with the winner determination problem.

For each auction, we know the identity of all participating firms and their characteristics as well as all the bids placed by them. In addition, we know the firm limits for the maximum number of TUs and market

share constraints in each auction. Finally, we know the set of winning bids in each auction and therefore, at every point in time, we know the identity of the firms serving each TU.

4 Estimation

This section describes details of the estimation of the structural model developed in Section 2 using the data from the Chilean School Meals auction that was described in Section 3. Our structural model identifies the costs of each firm based on equation (6). Similarly to CP, we use a two-stage method. In the first stage, described in Section 4.1, we parametrically estimate the distribution of competitors' bids, $H(\cdot)$, from the data. In the second stage, we estimate the winning probabilities of the package bids, $W^T G(b)$, and its Jacobian using simulation (see Section 4.2). We replace these values in equation (6) to obtain point estimates of the markups, and thereby the cost of each package. Section 4.3 provides the estimation results.

4.1 Estimating the Distribution of Competitors' Bids

As mentioned in Section 2.1, we conduct an auction-by-auction estimation of the distribution of competitors' bids. Given the high dimensionality of the bid vector b_f, which in our application is in the order of hundreds of packages, it is unfeasible to estimate the multivariate distribution of this bid vectors non-parametrically as was done by Guerre et al. (2000) in the single-unit auction case. Even in a small CA, Cantillon and Pesendorfer (2006) use a parametric approach. They assume a log-normal multivariate distribution, imposing constraints in the variance-covariance matrix. We follow a similar approach, as discussed next.

While the parametric approach provides a feasible estimation method in large-scale CAs, it is important to incorporate the needed flexibility so that the relevant factors that affect bid prices are captured in the model. First, in our application there is significant heterogeneity regarding the costs of serving different territorial units. These differences arise primarily because of location and density of schools; for example, units located in isolated rural areas tend to be more expensive than units in urban areas. Moreover, there may be substantial heterogeneity across firms' costs. For example, some firms have national presence, are vertically integrated, and have well functioning and efficient supply chains; while others are more rustic local firms. Hence, it is important to allow for sufficient flexibility in the distribution of bids to incorporate these two types of heterogeneity.

Second, package bids of the same bidder may be correlated. In our application, there are two main factors that can generate correlation between bids. First, a bidder that has a high cost in a given unit is likely to submit higher prices for all packages containing that unit. Second, if there are local advantages in the provision of services, a supplier with a low cost for a unit may also have low costs in nearby units. Hence, the unit composition of the package bids together with the spatial distribution of the territorial units provides a natural way to parameterize the covariance structure among package bids. We note that, as discussed in previous work by Olivares et al. (2011), the correlation structure of the competitors' bids has direct implications on the incentives to engage on strategic markup adjustments. Consequently, allowing for a flexible covariance structure that captures differences in the correlations between units can be important to identify the markups chosen by the different firms.

Third, bid data exhibits significant discounts both due to scale and density. Motivated by Olivares et al. (2011), we develop a parametric econometric model for package bids which captures these discounts, unit/firm heterogeneity and correlation between bids. Let v_i denote the volume of unit i, measured in million meals per year, and $v_a = \sum_{i \in a} v_i$ the total volume of package a. Bid prices are modeled as follows:[4]

$$b_{af} = -g^{scale}(v_a, \beta^{scale}_{s(f)}) - \sum_{c \in Cl(a)} g^{density}(v_c, \beta^{density}_{s(f)}) \cdot \frac{v_c}{v_a} + \sum_{i \in a} \tilde{\delta}_{if} \frac{v_i}{v_a} + \tilde{\varepsilon}_{af}. \tag{8}$$

The dependent variable, b_{af}, with some abuse of notation, denotes the *per-meal* price submitted by firm f for package a; that is, the actual bid price divided by the total volume of the package, v_a. The four terms in the right-hand side of equation (8) capture: (1) the effect of discounts due to scale (g^{scale}); (2) the effect of discounts due to density ($g^{density}$); (3) the effect of the specific units contained in the package; and (4) a Gaussian error term $\tilde{\varepsilon}_{af}$ capturing other factors affecting the bid price.

The distributional assumptions we make on the variables in equation (8) will induce distributions of competitors' bids $H(\cdot|Z_{f'})$ for the auction game of asymmetric information played by firms (see Assumption 4). An important distinction is which parameters of this equation are assumed to be known to the bidder at the time of choosing its bid, and which ones are unknown and thereby considered random from the bidders' perspective. These random parameters capture the asymmetric information among the bidders. This distinction between known and unknown parameters in equation (8) is important for simulating winning probabilities.

We use tilde (e.g. $\tilde{\delta}_{if}$) to denote factors that are *unknown* and random to the bidder. Hence, from the perspective of a bidder, model (8) is a regression with error components determined by $\{\tilde{\delta}_{if}\}_{i \in A}$ and $\{\tilde{\epsilon}_{af}\}_{a \in \mathcal{A}}$ and all the asymmetric information among bidders is encapsulated in those random components. In the context of the auction game, the parameters characterizing the discounts $(\beta^{scale}, \beta^{density})$ and the *distribution* of the error components $\{\tilde{\delta}_{if}\}$ and $\tilde{\epsilon}_{af}$ are considered common knowledge (as well as the bid data generating process specified by equation (8)). Therefore, to determine the distribution of competitors' bids as seen by the bidder, we need to estimate $(\beta^{scale}, \beta^{density})$ and the distribution of $\{\tilde{\delta}_{if}\}$ and $\tilde{\epsilon}_{af}$. Next, we discuss details of this estimation.

First, consider the terms capturing scale and density discounts, $(\beta^{scale}_{s(f)}, \beta^{density}_{s(f)})$. The model allows for some observed heterogeneity of these discounts across firms, with $s(f)$ indicating groups of firms of different business sizes. We found significant differences in discounts across the largest firms and the rest of the firms, so we grouped bidders into two groups, $s \in \{L, O\}$ (for Large and Other), to estimate discounts (L firms refer to the largest firms in JUNAEB's classification and can bid on packages of up to 8 units). There are several reasons that can help explain differences in discounts among firms of different size. First, bigger firms operate at a different scale and tend to operate other businesses outside the school meals procurement system. Hence, synergies for these firms could be different, which in turn could lead to different discounts. Note, however, that the discount functions g^{scale} and $g^{density}$ should not be interpreted directly as cost synergies because part of the discounts could arise from strategic behavior. In this regard, strategic markup adjustments could be different for firms that can bid on bigger packages, leading to further differences in the

[4]Olivares et al. (2011) report that when $\tilde{\delta}_{if}$'s are treated as parameters being estimated, the explanatory power of the regression is remarkably high with an R-square equals to 0.98.

discounts. We assume that the parameters $\{\beta^{scale}_{s(f)}, \beta^{density}_{s(f)}\}_{s(f)\in\{L,O\}}$ are common knowledge and that all the uncertainty associated to the magnitude of the discounts is provided by the error terms $\tilde{\epsilon}_{af}$.[5]

To measure scale discounts per meal, g^{scale} is specified as a step function of the package size v_a, and therefore the total discount in a package is a piece-wise linear function of the package size v_a. In contrast, density discounts depend on the proximity of the units in the package. To capture this, $g^{density}$ depends on the size of *clusters* of units in a package, where a cluster is a subset of the units in package a which are located in close proximity. In equation (8), $Cl(a)$ denotes the set of clusters in the package and c indicates a given cluster in this set, with size v_c. This approach follows directly from the work of Olivares et al. (2011) and further details on the computation of clusters are described in the appendix of that article.

Consider now the next-to-last term of the right-hand side of equation (8); this error-component term is a weighted average of the parameters $\tilde{\delta}_{if}$, which are firm and unit specific, capturing the effects of the individual units contained in package a. The $\tilde{\delta}_{if}$'s can be viewed as an individual price that bidder f is implicitly charging for unit i, net of any scale and density discounts. Note that this needs not be equal to the price the bidder charges for the stand-alone bid for unit i, because $\tilde{\delta}_{if}$ is an *average implicit price* considering all packages that contain the unit. These implicit prices could vary with the unit characteristics (e.g., urban vs. rural units) and local advantages of a firm in that unit, among other factors.

Note that we assume that the bidder does not observe the $\tilde{\delta}_{if}$'s of the competitors, which adds uncertainty to the competitor's bid distribution as observed by the bidder when choosing its bid. Instead, bidders view each competitor's $\tilde{\delta}_f = \{\tilde{\delta}_{if}\}_{i\in U}$ as a random vector drawn from a distribution which is common knowledge. Hence, we are not interested in point estimates of $\tilde{\delta}_f$'s per-se but rather the *distribution* of these average implicit prices as perceived by bidders. Accordingly, we let the vector of average implicit prices $\tilde{\delta}_f$ follow a multi-variate normal distribution and we seek to estimate the mean and covariance matrix of this distribution. Since our application has about 30 units on each auction, we need to impose further restrictions to estimate the covariance matrix. The specification we propose captures two important elements that are important for this application. First, some of the observed firms' characteristics, denoted by the vector Z_{if}, affect the implicit prices charged for the units. For example, Olivares et al. (2011) show that firms that seek to renew a contract they are already serving tend to offer more competitive prices. These firm characteristics are observed by all bidders and therefore considered common knowledge. Second, there is spatial correlation among the units, so prices for a unit tend to be positively correlated with the prices of nearby units. The following specification captures both effects:

$$\tilde{\delta}_{if} = \bar{\delta}_i + \beta^Z Z_{if} + \psi_{r(i),f} + \nu_{if}. \tag{9}$$

The parameters $\{\bar{\delta}_i\}_{i\in U}$ are treated as fixed-effects which capture the average implicit price for each unit charged among all bidders. Z_{if} denote the aforementioned unit/firm characteristics. The term $\psi_{r(i),f}$ is an error component associated with a pre-specified geographic region $r(i)$ where unit i is located; there are R different pre-specified regions. The error components $(\psi_{1f}, \cdots, \psi_{Rf})$ follow a multi-variate normal distribution with zero mean and covariance matrix Ω. Finally, the remaining error term ν_{if} follows an independent zero-mean normal distribution and is heteroskedastic with variance σ_i^2. These distributions are

[5] In small experiments we also found that modeling these discount parameters as random variables does not affect the cost estimates by much.

common knowledge. Because R may be much smaller than the number of units, this specification provides a substantial dimensionality reduction over the fully flexible distribution of $\tilde{\delta}_{if}$'s.

Under the specification (9), the covariance structure of any two average implicit prices $\tilde{\delta}_{if}$ and $\tilde{\delta}_{jf}$ is given by: $\text{Cov}(\tilde{\delta}_{if}, \tilde{\delta}_{jf}) = \Omega_{r(i),r(j)} + \sigma_i \sigma_j \mathbf{1}[i = j]$. Thus, under this model two unit prices will be more positively correlated if the regional effects of the corresponding regions are more positively correlated. Note that this specification imposes positive correlation among unit prices in the same region; this pattern is observed in the data. However, it is flexible in allowing positive or negative correlation among units in different regions.

In summary, the competitors' bid distribution $H(\cdot|Z_f; \phi)$, where ϕ is the set of distribution parameters, is a mixture distribution described by the following terms: (1) the deterministic component associated with the discounts g^{scale} and $g^{density}$, captured by the vector parameters β^{scale} and $\beta^{density}$; (2) a random component associated with the average implicit price vector $\tilde{\delta}_f$, whose distribution is fully described by the vector parameters $\bar{\delta} = (\bar{\delta}_1, \cdots, \bar{\delta}_N)$, β^Z, $\sigma = (\sigma_1, \cdots, \sigma_N)$ and the covariance matrix Ω; and (3) a Gaussian error component $\tilde{\varepsilon}_{af}$, which we assume to be heteroskedastic so that the variance depends on the number of units in the package, $\sigma^{\varepsilon}_{|a|}$. All of these parameters that characterize the distribution of competitors bids are considered common knowledge and need to be estimated from bidding data, so that we can use the competitors' bid model in the simulation of the winning probabilities.

The following two-step procedure is used to estimate the econometric model defined by equations (8) and (9):

- First step: estimate (8) via a Generalized Least Squares (GLS) regression to obtain estimates of β^{scale}, $\beta^{density}$, and point estimates of the implicit prices $\tilde{\delta}_{if}$'s.
- Second step: plug-in the estimated $\tilde{\delta}_{if}$'s into equation (9) and estimate its model parameters through maximum likelihood.

The identification of model (8) is based on variation across package bids in a single auction, and hence requires a large number of package bids. More specifically, the two step procedure described above estimates scale and density discounts using variation across different combinations submitted by the same firm over the same set of units. Under the usual orthogonality conditions of GLS, the first step regression provides consistent estimates of β^{scale}, $\beta^{density}$ and point estimates of the $\tilde{\delta}_{if}$'s.[6]

Identification of the parameters in model (9) is based on variation across units and firms. Given consistent estimates of the implicit unit prices $\tilde{\delta}_{if}$, the second step provides consistent estimates of $\{\bar{\delta}_i, \sigma_i\}_{i \in U}$, β^Z and Ω as long as Z_{if} is orthogonal to the error components $\psi_{r(i),f}$ and ν_{if}. The consistency of our two-step method is a special case of the 2-step M-estimators described in Wooldridge (2002).

Estimates for the Parameters of Bid Distribution. We provide the results for the 2003 auction. Table 1 reports estimates of β^{scale} and $\beta^{density}$ from the first step regression. The scale and density per-meal

[6]Point-identification of the implicit unit prices $\tilde{\delta}_{if}$ can be obtained when each firm submits many bids containing each unit i. Our empirical application meet these requirements. For example, in 2003, 20 firms participated for 33 units and on average submitted around 2100 bids each.

discount curves, $g^{scale}(v_a, \beta^{scale}_{s(f)})$ and $g^{density}(v_c, \beta^{density}_{s(f)})$, are specified as step functions with interval of three million meals per year in the package volume v_a and cluster size v_c, respectively. Each number indicates the average discount in per-meal price when units are combined to form a package that belongs to the corresponding volume level. For example, when units are combined into package a with volume $v_a \in [18, 21]$, then on average a large firm submits a bid that is Ch$ 22.78 cheaper per meal than the weighted average bid price of those individual units in the package. If all these units are located nearby and form a cluster, there is an additional discount of Ch$11.27 on average for a large firm. The results show that large firms were able to provide higher discounts which amounts up to 9% of average bid price. All the coefficients are estimated with precision at the significance level of 0.01%.

In addition, the first-step estimation of regression (8) provides point estimates of the implicit average prices δ_{if}'s (not shown). On average, the standard error for these estimates are in the order of 0.5% of the point estimates, which is reasonably accurate. To further validate these estimates we compared them with the stand-alone bids b_{if}'s. The ratio b_{if}/δ_{if} is on average 0.998 with standard deviation of 0.026. The correlation of the two measures is 0.987. These results suggest that the implicit average prices effectively separate out the individual prices from package discounts.

The second step estimation of regression (9) provides estimates for the distribution of the average implicit prices δ_{if}'s, characterized by $\{\bar{\delta}_i, \sigma_i\}_{i \in U}$, the covariance matrix Ω and β^Z, the coefficients of the firm characteristics. Firm characteristics include an indicator on whether the firm was awarded the unit in the previous auction (we also tried other firm characteristics but those were not statistically significant). Due to space limitations, we do not report the estimates of the $\bar{\delta}_i$ parameters, but these were estimated with precision - on average, the standard errors are 1.2% of the point estimates. The estimated coefficient for the incumbency effect (β^Z) is -6 with a p-value of 0.012, suggesting that on average incumbent firms submit bids that are around 1.5% cheaper than non-incumbent firms.

Table 2 shows the correlations between the region effects $\psi_{r(i),f}$ (which were calculated based on estimates of the variance/covariance matrix Ω). These estimates imply a significant positive correlation among units: on average, the correlation between the implicit prices of two units in the same region is 0.68, and 0.45 for units located in different regions. The last column of the table shows the standard deviations of each region effect $\psi_{r(i),f}$ (which corresponds to $\sqrt{\Omega_{rr}}$ for each region r). All the standard errors of the second step estimates are computed via a parametric bootstrapping procedure.

4.2 Markup and Cost Estimation

Using the estimated distribution of competitors' bids, markups are estimated using equation (6) for each firm. This requires calculating the aggregated winning probabilities $W^T G(b)$ and its Jacobian, as described in this section.

First, the specification of the package-characteristic matrix W is based on a refinement of the sized-based markup approach described in section 2.3. Package size is measured in terms of the number of units in the combination, so that the packages with same number of units are grouped together. Each size group is then partitioned into those that should have similar markup levels in the full-dimension model, based on the heuristic described in section 2.3.

In our application, package volume – defined as $v_a = \sum_{i \in a} v_i$ – has a first-order effect on the bid price and so prices may vary substantially even within each size group. For this reason, we assume that bids within each group have a common *per-meal* markup, instead of a fixed *absolute* markup. Defining b_a and c_a as the per-meal bid and per-meal cost of package a, and defining the non-zero components of W as $W_{as} = v_a$, the firm's decision variable θ can be interpreted as a per-meal markup vector.[7]

Next, we describe a Monte Carlo simulation to calculate the winning probabilities for a given bidder f. Given the point estimates $\hat{\phi}$ for the distribution parameters from our two step approach, each simulation run l consists of the following:

1. For each competitor f' (different from bidder f), draw independently a bid vector $b_{f'}^{(l)}$ (containing all submitted packages by that firm) from the estimated bid distribution $H(\cdot|Z_{f'}; \hat{\phi})$.

2. Using the observed bid b_f for bidder f and the sampled competitors' bids $\{b_{f'}^{(l)}\}_{f' \neq f}$, solve the winner determination problem.

3. Let $\iota^{(l)}$ be a vector of A binary variables indicating the packages awarded to the bidder f. Store in memory the vector $W^T \iota^{(l)}$.

At the end of the simulation after L replications, the aggregated winning probability vector can be estimated by:

$$W^T G(b) \approx \frac{1}{L} \sum_{l=1}^{L} W^T \iota^{(l)}.$$

Note that if the distribution of competitors' bids is estimated consistently, then the previous equation provides consistent estimates of the aggregated winning probabilities as L becomes large. Finite differences are used to compute the Jacobian of $W^T G$, which requires calculating the change in the winning probabilities from a small change in each markup variable θ_i. Because the bid is linear in θ (by Assumption 5), this is equivalent to consider a small change in the observed bid vector b in the direction of that markup variable. Specifically, consider a change in the j^{th} component of the markup vector, and let W_i be the i^{th} column of the package-characteristic matrix W. To calculate the i^{th} row and j^{th} column element of the Jacobian we use the central finite difference method:

$$[\mathcal{D}_\theta W^T G(b)]_{ij} = \frac{\partial W_i^T G(b)}{\partial \theta_j} \approx \frac{W_i^T G(b + hW_j) - W_i^T G(b - hW_j)}{2h}.$$

The computation of $W_i^T G(b + hW_j)$ and $W_i^T G(b - hW_j)$ is done via simulation as before: on each simulation run, we solve the winner determination problems with the perturbed bid prices and keep track of the winning bids. Once the aggregated winning probability vector $W^T G(b)$ and its Jacobian matrix $\mathcal{D}_\theta W^T G(b)$ are estimated, the markup vector θ for this bidder is obtained through the identification equation (6): $\theta = -\left\{[\mathcal{D}_\theta W^T G(b)]^T\right\}^{-1} W^T G(b)$. The winning records used in the central finite difference method

[7] Recall that in Proposition 2 we assumed that $\{\theta_a\}_{a \in \{1,\ldots,K\}}$ and θ_u are absolute markups. In the per-meal markup specification, we obtain absolute markups by multiplying them by the total meal volume in the package through the W matrix. In this case, the α_a's are not necessarily all negative a priori. However, in our estimations they turned out to be negative for most of firms and auctions, leading to weighted average markups when aggregated.

also enable to estimate the direct second-order derivatives of the bidders' expected profits with respect to each of their markup variables. We obtained that these estimates are negative for all firms, which is consistent with the local optimality of the estimated markups. [8]

4.3 Estimation Results

The estimated markups reveal reasonable levels for most of firms with average margin of 2.7%. However, from the 20 participating firms there are two extreme firms whose markups are unreasonably high and lead to negative costs for some packages. This indicates that the assumptions on the bidders behavior may not be satisfied for these firms. Because we are unable to correctly infer the cost information of these firms and because these firms did not win any contracts in the 2003 auction, we omit them from our analysis hereafter.

We observe roughly three groups of firms. The first group, which we call "aggressive" group, consists of nine firms whose total winning probabilities aggregated over all packages are higher than 50%. The other firms have very low winning probabilities (less than 2%) except for two firms whose total probabilities are 44% and 14%. In terms of markups, the aggressive firms set markups ranging from 2% to 18% of the average bid price with an average markup of 4.7% of the average bid price (US$ 0.88 per meal). As expected, the other firms set lower markups, resulting in an average markup overall firms of around 2.8% of the average bid price. Table 3 shows the average per-meal markup estimates of each package size for three representative firms of different level of "aggressiveness". In addition, the estimates indicate that firms reduce their markups as the size of packages increases, showing that some portion of the discounts in package bids are due to markup adjustments.

Firms submit hundreds to thousands of bids, and about 13% of them are relevant bids. However, for the aggressive group, this increases to 22%. In addition, as we have markup and cost information of relevant bids, we are able to compute the total cost and markup of the CA allocation. The total procurement cost was US$ 70.5 million per year and the supplying costs was US$ 67.2 million per year, which yields 4.8% of average profit margins to winning firms. This level of profit margins is consistent with the Chilean government's estimate for this market, which is re-assuring.

Finally, as a robustness check, we also performed the estimation for the 2005 auction, where 16 firms participated for 23 units. The results look consistent with the 2003 auction, both in the shape and level of the estimated markups. The total procurement cost amounts to US$ 53.4 million and the total supplying cost is US$ 51.5, which give 3.5% of average profit margins to winning firms.[9]

[8] We run 100,000 simulation runs for each firm. On average, it takes around 15 days to finish the simulation for the firms with relatively higher winning probabilities with about 10 - 18 markup variables. For other firms, that have relatively lower winning probabilities, it took around 10 days with about 5 markup variables. The program is implemented in C with CPLEX V12.1 and ran on Columbia's research grid where each machine has eight 2.4GHz CPUs. Note that to fully evaluate the local optimality of the markups, we need to estimate the Hessian matrices of the bidders' expected profit. However, this is computationally very intense, requiring an order of magnitude more computation time to what is required to estimate the markups.

[9] In 2004, the government introduced an electronic bidding system to the auction process that resulted in a huge increase in the number of submitted bids. On average, firms placed four times as many bids as they did in 2003. Moreover, the number of firms and auctioned units were also larger, and we omit the results of this year as it requires an onerous amount of computational time. However, the estimation was more manageable for the 2005 auction as the number of units auctioned and the participating firms are smaller. We did not estimate years 1999-2002, because in those auctions bidders had less experience and history to rely on, and

Next, we evaluate the cost synergies – cost savings from combining units together – implied by the estimates. As mentioned in Section 3, there could be two important sources of synergies in our application: i) synergies due to economies of scale, which depend on the total volume of the package that is supplied; and ii) synergies due to economies of density achieved when nearby units are supplied together. Next, we describe how to calculate both type of synergies based on the estimated markups.

The per-meal cost of each package a submitted by firm f is given by $c_{af} = b_{af} - w_a\theta_f/v_a$, where θ_f is the markup vector estimated for that firm, b_{af} is again the per-meal bid price placed by firm f for package a, and w_a is the a^{th} row of package-characteristic matrix W used for bidder f. The per-meal cost synergy in this package, denoted by s_a, can then be calculated as $s_a = \sum_{i\in a} \frac{v_i}{v_a} c_i - c_a$, where c_i is the point estimate for the cost of unit i. Table 4 shows some summary statistics of the cost synergies. The results of this calculation suggest there are significant cost synergies amounting up to around 5% of the average bid price.

One disadvantage of the synergies that are shown in Table 4 is that they rely only on the sample of relevant bids, with a few firms accounting for a disproportionate fraction of this sample. This is because this direct calculation of cost synergies requires the costs of single-unit packages, and many of these single-unit bids are irrelevant. In order to use a larger portion of the packages to estimate cost synergies, we run a regression similar to (8) but where the dependent variable is the estimated per-meal cost of the package (rather than the per-meal bid). Here, we also try to disentangle how much of these cost synergies arise from economies of scale and density:

$$c_{af} = \sum_{i\in a} \xi_{if} \frac{v_i}{v_a} - g^{scale}(v_a, \gamma^{scale}_{s(f)}) - \sum_{c\in Cl(a)} g^{density}(v_c, \gamma^{density}_{s(f)}) \cdot \frac{v_c}{v_a} + \varepsilon_{af}, \tag{10}$$

where again $s(f) \in \{L, O\}$ indicates one of the two firm group sizes. As in (8), g^{scale} and $g^{density}$ are specified as step functions of the size of the package (v_a) and the cluster (v_c), respectively. The parameters ξ_{if} represent an *implicit cost* of each unit, which is estimated for all units, including those units for which the single-unit package was irrelevant. Note that for this regression we use all relevant packages. To validate this approach, we compared the estimates of these implicit costs with those estimated directly via the structural estimation (over the sample of relevant single-unit bids). The correlation between the two is 0.9931 and the average absolute difference is 1.04% of average unit cost. The ratio c_{if}/ξ_{if} averages to 1.000 with standard deviation of 0.011. Hence, equation (10) seems a reasonable approach to estimate cost synergies.

Figure 2 shows the estimated cost synergies (from equation (10)) together with the bid discounts estimated previously from equation (8) over the sample of relevant bids. The results are shown for each of the firm groups.[10] The results show that while there is some strategic markup adjustments, most bid discounts are actually explained by cost synergies. These synergies are quite significant and can be as large as 6% of the bid price on average. The results also show that economies of scale are predominant, but that economies of density are also important and can account for 1% of the average bid price.

Finally, similar to equation (8) and equation (10), regression equations without the density terms

were less sophisticated, so that our structural model assumptions may be harder to justify.

[10]There are two small firms whose estimated cost synergies are significantly different from the rest firms, and they are not accounted in the figure.

$g^{density}$ will capture the average discount and synergy levels in terms of size of the packages. This information is also useful because it gives us the *overall* level of bid discounts and cost synergies due to the combined effect of economies of scale and density. Figure 3 shows the estimated overall bid discounts and cost synergies of those 16 firms using the relevant bids and the corresponding costs. The results suggest that on average, most of the bid discounts (at least 75%) are driven by cost synergies as oppose to strategic markup adjustments.

5 Efficiency Analysis and Counterfactuals

The previous results seem to suggest that in our application allowing package bidding may be appropriate: cost synergies are significant and account for most bid discounts vis-á-vis strategic markup adjustments. Moreover, the overall markups that firms gain do not seem too large, resulting in a reasonable total procurement cost. In addition, the results suggest that the bidding language should allow bidders to express both economies of scale and density. Overall, our results suggest that the advantages of using package bidding (allow bidders to express cost synergies) may be larger than its disadvantages (the additional flexibility that firms can use to strategize and game the mechanism).

While suggestive, the previous results are not conclusive. In this section we use our estimates to provide sharper results concerning the efficiency and procurement cost of our CA. In particular, we study the allocative efficiency and procurement cost of the first-price sealed-bid CA, and compare it to alternative auction mechanisms. We provide results for the 2003 auction.[11]

5.1 Performance of the First-Price CA

In this section we study the allocative efficiency of the first-price CA. The winning bidders' costs under the first-price CA allocation can be directly computed using the cost estimates obtained in Section 4.2. If we had the cost estimates for *all* possible unit combinations, one could also calculate the minimum-cost allocation. Unfortunately, the structural estimation only identifies the costs of relevant bids, and the minimum-cost allocation over this subset of combinations could overestimate the true minimum-cost allocation that also considers irrelevant bid combinations.

To address this issue, we propose estimating the cost of irrelevant bid packages through an out-of-sample extrapolation based on equation (10). However, the total number of feasible packages are in the order of millions and it is computationally infeasible to extrapolate to the entire set of (out-of-sample) packages. Instead, we choose the set of packages on which at least one bidder placed a bid, which is in the order of 30 thousand packages. We call this the *expanded package set*. Then for each firm, we extrapolate costs to this expanded package set as long as the package satisfies the maximum TU and global market share constraints for that firm. While this is a small subset of the entire packages, it provides a reasonable approach to extending the set of bids observed in the data.

[11] The results for the 2005 auction are similar and consistent with the 2003 auction. We provide the counterfactual results for the 2005 auction in Online Appendix D.

This approach implicitly assumes that the selection of the bids in the irrelevant bid sample is independent of the costs of these units. Recall that irrelevant bids include bids that were not submitted by the bidder. Hence, in our application, it could be possible that the sample selection of irrelevant bids is related to costs: for example, bidders are likely to bid on the subset of combinations where they are more competitive, so that higher-cost combinations are not submitted. If this is the case, then our cost extrapolation procedure could lead to a minimum-cost allocation which is lower than the true one, so that we could *overestimate* the true efficiency-loss of the first-price CA.

Recall that in 2003, the bidders' supplying costs given by the auction allocation were equal to US$ 67.2 million per year. The cost efficient allocation that minimizes costs over the set of relevant bids and feasible allocations (considering the constraints described in Section 3) is equal to US$ 66.7 million per year. If we also consider the cost extrapolation to irrelevant bids as described above (i.e., over the expanded package set), the minimum-cost allocation goes down to US$ 66.2 million per year. This implies an efficiency loss of the first-price CA of 1.5%, which is evidently low.[12] The high efficiency and relatively small profit margins for firms (around 5% as presented in Section 4.3) achieved by the school meals CA suggests that it is a reasonable mechanism for the procurement of this public service.

5.2 The VCG Mechanism

While our previous results supports that using the first-price CA in our application seems appropriate, it is also useful to compare the performance with alternative auction mechanisms. Doing these counterfactuals on alternative mechanisms requires computing the bidding strategies played in equilibrium by the bidders. Unfortunately, there are few equilibrium results for most of the multi-unit auction mechanisms that are used in practice. For example, it is intractable to compute equilibria of the respective games of asymmetric information for our CA, for independent single-unit auctions under the presence of synergies, or for the multi-round CAs used by the FCC in the wireless spectrum auctions.

We can perform, however, a counterfactual with the Vickrey-Clarke-Groves (VCG) mechanism, which is a generalization of a second price auction: the payment to a winner is essentially the cost of providing the units she wins in the lowest cost allocation without her, and loosing bidders do not receive payments. It is well known that under this payment rule, truth-telling is a dominant strategy, i.e., bidders report their true costs. In Online Appendix C we provide details about VCG and its payment rule. Like in the first price CA, winners in VCG are also determined by finding the combination bids that achieve the minimum procurement cost; this results in the efficient allocation due to the truth-telling property. However, despite these theoretical virtues, VCG mechanisms have been criticized for other numerous drawbacks, leading to a very rare use in practice. In particular, Ausubel and Milgrom (2006) have shown that in the face of complementarities, the VCG procurement costs can be prohibitively high. This and other deficiencies of VCG in settings with complementarities have motivated an active research agenda in recent years that studies alternative payment

[12] It is worth noting that the first-price CA tends to identify the most cost efficient firms in the different geographical regions. More specifically, there are nine firms in the CA allocation and ten firms in the efficient allocation; the majority of them –seven firms– are present in both cases. Two firms are allocated the exact same set of packages in both cases and other firms win packages that contain many overlapping units or units from the same geographical regions.

rules, giving rise to the so-called "core-selecting auctions" (we provide more details below). Hence, it is on itself interesting to see how VCG mechanisms would perform in real-world applications.

In our analysis we use the same set of extrapolated bids as in Section 5.1 as the bids (costs) that bidders would report in a VCG mechanism. We know the VCG allocation is efficient, so it coincides with the minimum-cost allocation (that satisfies all constraints described in Section 3.2); this was previously computed in Section 5.1.From the bids, we can compute the individual VCG payments to the winning bidders (see online appendix), and by summing them, we obtain the VCG procurement cost. As seen in the previous section, the total annual procurement cost in the 2003 first-price CA is US$ 70.5 million. The total annual procurement cost under the VCG mechanism is US$ 70.3 million, which is about 0.32% cheaper than the first-price CA.

Given the theoretical literature mentioned above describing the pitfalls of VCG, the result is striking; in our application, VCG achieves payments comparable to the first-price CA, so in fact VCG induces reasonable procurement costs. We believe this result is driven by the significant amount of competition introduced by the large number of package bids submitted by firms. In this case, a winning bidder is not that relevant; if her bids are eliminated, there is another allocation that achieves costs close to the minimum-cost allocation, leading to reasonably low VCG payments. More broadly, it is interesting to note that in the examples provided by Ausubel and Milgrom (2006), the amount of competition is limited, resulting in high VCG payments. We believe that VCG should achieve reasonable procurement costs in settings with a reasonable amount of bidders that are able to submit many package bids. The latter should be expected when it is relatively effortless for a bidder to evaluate its cost in an additional package, therefore, the bidder can easily submit many package bids.

We finish this subsection by observing that Ausubel and Milgrom (2006) and Milgrom (2004) show that the poor performance of VCG arises from the fact that the VCG outcome may not be in the *core* of the transferable utility cooperative game played among the bidders and the buyer (auctioneer). In this sense, the core can be understood as a *competitive* benchmark; if the outcome is not in the core, the payments are so high that there is a group of bidders that can offer a more favorable deal to the auctioneer. In our application, we find that indeed the VCG payoffs lie *essentially* in the core, which is consistent with the reasonable total procurement cost achieved by VCG. In particular, following Day and Raghavan (2007) we find the closest point in the core (with respect to the truthful bids) to the VCG payments under a suitable norm. We find that the differences in total procurement costs between these two points is only 0.1% in 2003. Also, individual payments are very similar as well; half of the winners receive exactly the same payments in the core point as in VCG, and the rest receive payments that are no more than 0.7% apart. We provide more details about the core analysis in Online Appendix C.

5.3 Supplier Diversification

The CA of our application imposes three types of constraints aimed at preserving a more diversified supplier base: (1) a single bidder cannot be awarded more than 16% of the total volume including outstanding contracts awarded in previous years (*market share* constraint); (2) a minimum number of winning firms on each auction (*global competition* constraint); and (3) a minimum number of winning firms on each of the 13

pre-specified geographic regions (*local* constraints). We now focus on measuring what is the efficiency loss imposed by these constraints.

To study efficiency of the first-price CA, we have already calculated the minimum-cost allocation that satisfies these constraints. We could compare this with the minimum-cost allocation obtained under the larger feasible set of allocations when the constraints are removed. However, this may not be a fair comparison because bids on packages that violate some of the constraints are not submitted by the bidders. In other words, in the counterfactual world without the constraints we should observe new package bids that are not observed under the current format with the constraints. To address this issue, we expand the set of submitted bids in the counterfactual without the constraints as we now explain.

First, consider the market share constraint that imposes a maximum volume of 16% of the total volume of the country, equivalent to about $K = 40$ million meals per year to each firm. Under this constraint, bids on packages with larger volume than K will never be observed in the data. It turns out that because of the 8 unit limit for the packages, the market share constraint is never binding for those firms which do not have any existing outstanding contracts, because the maximum volume that can be achieved with 8 units is less than K. So those firms do place bids on packages of any volume with at most 8 units. We call such firms whose bidding is not limited by the market share constraint the *unrestricted firms*.

To extrapolate costs to packages violating the market share constraint we do the following. Consider a large bidder f that has existing outstanding contracts for a total volume of X. This firm can only submit packages of volume less than or equal to $K - X$. Removing the constraint would allow this bidder to submit packages of any volume up to K, as long as they have 8 units or less. We denote by A_X the set of observed combinations that are infeasible for bidder f but feasible for the unrestricted bidders, hence contained in the expanded package set. We can use regression (10) to predict the costs of combinations in A_X for bidder f. Doing this for all bidders allows us to build a larger feasible set that contains bids that would not be feasible when the 16% market share constraint is included. Again the expanded package set which is in the order of 30 thousand packages – still less than the 20 million possible packages that could be submitted – and provides a reasonable set of bids to evaluate the effect of removing the market share constraints.

In contrast, packages that violate the local competition constraints are almost never observed. To illustrate why this is the case, consider region 13 which has seven units but the minimum number of firms required to win is four. Hence bids on any package containing five of more units in region 13 will violate this constraint and will never win. Note that unlike the market share constraint which is a firm-wise restriction, local constraints are applied to all firms and hence no such packages are found in the expanded package set. For this reason, we cannot analyze the effect of removing the local constraints using the same approach to expand the set of bids. Finally, we note that it is not a priori clear whether removing the global competition constraint would result in significantly different package bids submitted, because in any event firms cannot submit packages larger than 8 units. Therefore, we do not include additional bids associated to removing that constraint, and we focus on the efficiency loss caused by the market share constraint and the global competition constraint.

To measure the efficiency loss due to the market share and the global competition constraints, we compare the minimum-cost allocations with all those constraints and without the two types of constraints.

We find that this efficiency loss in 2003 is about 0.57%, which is relatively small.[13] The small impact of these constraints on efficiency can be partially explained by the structure of the cost synergies in the industry. As we saw in section 4.2, scale cost synergies get exhausted, so there are small cost reductions for combinations that lie beyond the volume range that is currently feasible in the auction. To further evaluate the inclusion of these constraints in the mechanism, it would be useful to measure the value that the constraints aimed at promoting supplier diversification bring in terms of increased competition. We leave this analysis for future research.[14]

6 Conclusions

In this paper we develop a structural estimation approach for large-scale first-price CAs and applied it to the Chilean school meals auction. An important methodological contribution in our work is to introduce a reduced dimensional markup model in which bidders are assumed to determine their markups based on a reduced set of package characteristics. Our modeling approach is essential to achieve computational tractability for estimation in a large-scale CA.

We find that cost synergies in the Chilean school meals auction are significant and the current CA mechanism, which allows firms to express these synergies through package bidding, seems appropriate. In particular, the current CA achieves high allocative efficiency and a reasonable procurement cost. We also find that the effect of some of the side constraints currently used in the CA, which limit the market share each bidder can get in order to promote suppliers' diversification, results in only a small efficiency loss. We also compared the performance of the VCG mechanism to the first price CA used on this application. Contrary to results obtained in previous theoretical work where VCG has been criticized for achieving high procurement costs, we find that the total VCG payment is reasonable and quite close to the first-price CA payment.

Overall our results provide useful insights for the design of the Chilean auction. More broadly, our results highlight that the simultaneous consideration of the firms' operational cost structure and their strategic behavior is key to the successful design of a CA. Moreover, our structural estimation framework is sufficiently general to be used in other applications of large-scale CAs. In this way, we hope that this research agenda enhances the understanding of the performance of CAs and thereby provide insights to improve their design.

[13] The final allocations in both cases look similar. Nine firms win in both cases and only one winner is replaced by another. Two firms win exactly the same packages, and six other firms have many of the winning units overlap in both cases or win units in the same region. The efficiency loss is mainly triggered by one large firm who won a package of two units with market share constraint and won a package of five units in the unconstrained case.

[14] Olivares et al. (2011) show that local competition, measured by the number of firms serving nearby units, has a significant effect in reducing prices in this application. This suggests that supplier diversification at a local level can lead to increased competition.

References

Aksin-Karaesmen, Z., B. Ata, S. Emadi, C. Su. 2011. Structural estimation of callers' delay sensitivity in call centers. Working Paper.

Allon, G., A. Federgruen, M. Pierson. 2011. How much is a reduction of your customers' wait worth? an empirical study of the fast-food drive-thru industry based on structural estimation methods. *Manufacturing & Service Operations Management* **13**(4) 489 – 507.

Athey, S., P. A. Haile. 2006. Empirical models of auctions. Invited lecture for the Ninth World Congress of the Econometric Society.

Ausubel, L. M., P. Cramton, R. P. Mcafee, J. Mcmillan. 1997. Synergies in wireless telephony: Evidence from the broadband pcs auctions. *Journal of Economics and Management Strategy* **6**(3) 497–527.

Ausubel, L. M., P. Milgrom. 2006. The lovely but lonely Vickrey auction. P. Cramton, Y. Shoham, R. Steinberg, eds., *Combinatorial Auctions*. MIT Press, 17–40.

Cantillon, E., M. Pesendorfer. 2006. Combination bidding in multi-unit auctions. Working Paper.

Chapman, J.T.E., D. McAdams, H.J. Paarsch. 2005. Multi-unit, sealed-bid, discriminatory-price auctions. Working Paper, University of Iowa.

Cramton, P., Y. Shoham, R. Steinberg. 2006. *Combinatorial Auctions*. MIT Press.

Day, R., P. Milgrom. 2008. Core-selecting package auctions. *International Journal of Game Theory* **36**(3–4) 393–407.

Day, R., S. Raghavan. 2007. Fair payments for efficient allocations in public sector combinatorial auctions. *Management Science* **53**(9) 1389–1406.

Epstein, R., L. Henríquez, J. Catalán, G. Y. Weintraub, C. Martínez. 2002. A combinational auction improves school meals in Chile. *Interfaces* **32**(6) 1 – 14.

Fox, J. T., P. Bajari. 2011. Measuring the efficiency of an FCC spectrum auction. Working Paper.

Gandal, N. 1997. Sequential auctions of interdependent objects: Israeli cable television licenses. *The Journal of Industrial Economics* **XLV**(3) 227–244.

Guerre, I., I. Perrigne, Q. Vuong. 2000. Optimal nonparametric estimation of first-price auctions. *Econometrica* **68**(3) 525–574.

Hendricks, K., R. H. Porter. 2007. An empirical perspective on auctions. M. Armstrong, R. Porter, eds., *Handbook of Industrial Organization, Volume 3*. Elsevier.

Horn, R. A., C. R. Johnson. 1985. *Matrix Analysis*. Cambridge University Press.

Hortaçsu, A., D. McAdams. 2010. Mechanism choice and strategic bidding in divisible good auctions: An empirical analysis of the turkish treasury auction market. *Journal of Political Economy* **188**(5) 833–865.

Kastl, J. 2011. Discrete bids and empirical inference in divisible good auctions. *Review of Economic Studies* **78**(3) 974–1014.

Li, J., N. Granados, S. Netessine. 2011. Are consumers strategic? structural estimation from the air-travel industry. Working Paper.

Milgrom, P. 2000. Putting auction theory to work: The simultaneous ascending auction. *Journal of Political Economy* **108**(2) 245 – 272.

Milgrom, P. 2004. *Putting Auction Theory to Work*. Cambridge.

Moreton, P. S., P. T. Spiller. 1998. What's in the air: Interlicense synergies in the federal communications commission's broadband personal communication service spectrum auctions. *The Journal of Law and Economics* **XLI** 677–716.

Olivares, M., C. Terwiesch, L. Cassorla. 2008. Structural estimation of the newsvendor model: An application to reserving operating room time. *Management Science* **54**(1) 41 – 55.

Olivares, M., G. Y. Weintraub, R. Epstein, D. Yung. 2011. Combinatorial auctions for procurement: An empirical study of the chilean school meals auction. Forthcoming, *Management Science*.

Paarsch, H. J., H. Hong. 2006. *An Introduction to the Structural Econometrics of Auction Data*. MIT Press.

Reguant, M. 2011. The welfare effects of complementary bidding mechanisms: An application to electricity markets. Working Paper.

Wooldridge, J.M. 2002. *Econometric analysis of cross section and panel data*. The MIT press.

Tables and Figures

Large Firms			Other Firms		
Volume	Scale	Density	Volume	Scale	Density
$[3, 6]$	8.33 (1.30)	6.46 (0.51)	$[3, 6]$	8.50 (0.62)	1.82 (0.14)
$[6, 9]$	15.21 (1.33)	7.81 (0.53)	$[6, 9]$	11.86 (0.64)	3.31 (0.19)
$[9, 12]$	17.82 (1.31)	8.10 (0.55)	$[9, 12]$	13.50 (0.65)	3.92 (0.24)
$[12, 15]$	19.10 (1.30)	8.57 (0.56)	$[12, 15]$	13.44 (0.67)	5.69 (0.28)
$[15, 18]$	20.76 (1.29)	9.13 (0.57)	$[15, 18]$	12.42 (0.69)	6.96 (0.36)
$[18, 21]$	22.78 (1.30)	11.27 (0.65)	$[18, 21]$	10.90 (0.72)	
$[21, 24]$	24.38 (1.30)				
$[24, 27]$	24.95 (1.35)				

Table 1 – Results from the first step regression (equation (8)) for 2003 auction. Robust standard errors are shown in parenthesis. Combination/Cluster volume is measured in million meals per year.

	Correlation Coefficients					Std. Dev.
Region	4	5	9	12	13	
4	1.00 (0.00)	0.52 (0.21)	0.31 (0.27)	0.45 (0.24)	0.67 (0.17)	14.56 (3.20)
5	-	1.00 (0.00)	0.65 (0.16)	0.69 (0.17)	0.69 (0.13)	14.52 (2.55)
9	-	-	1.00 (0.00)	0.42 (0.22)	0.09 (0.27)	22.92 (4.02)
12	-	-	-	1.00 (0.00)	0.48 (0.22)	46.48 (9.97)
13	-	-	-	-	1.00 (0.00)	13.46 (2.29)

Table 2 – Results from the second step regression (equation (9)) for 2003 auction. Standard errors are shown in parenthesis. Standard deviations of regional effects are measured in Chilean Pesos.

Firm	Prob	1	2	3	4	5	6	Average
47	0.9193	22.64	15.07	12.14	7.98	7.54	7.19	9.88
36	0.6642	3.00	2.39	2.21	1.77	1.50	1.41	2.07
19	0.1578	0.81	0.82	0.84	0.79	0.72	0.71	0.79

Table 3 – Results from the markup estimation for representative firms of different winning probability levels for 2003 auction. Prob refers to the probability that the firm wins any package. The rest are the average per-meal markups corresponding to each package size. The markups are shown in the percentage of the average bid price per meal (US$ 0.88).

Size	2	3	4	5	6	7	8	Average
Cost Synergy (CH $)	5.13	10.64	14.40	13.71	15.07	16.64	18.93	12.82
% of Average Bid Price	1.28	2.66	3.60	3.43	3.77	4.16	4.73	3.21
Number of Observations	289	87	121	49	126	169	205	

Table 4 – Average cost synergies computed directly from estimated costs of individual units and multi-unit packages for 2003 auction. Size refers to the number of units in the package. The cost synergy measures the average per-meal cost savings when the units are combined to form a package of given size.

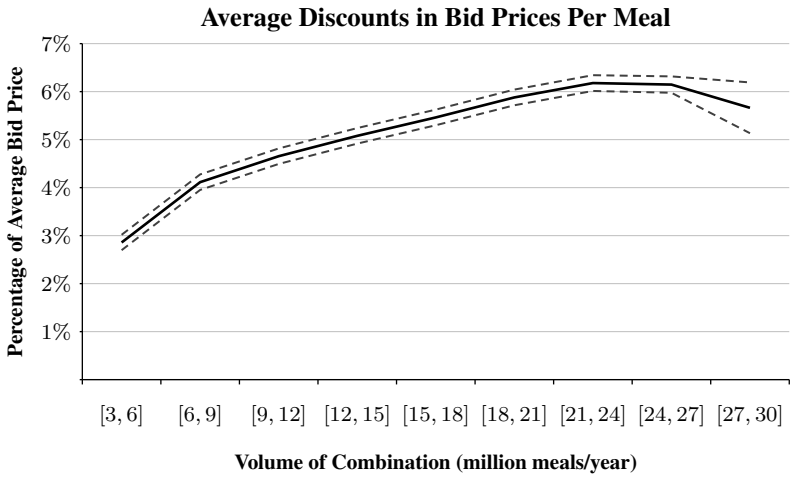


Figure 1 – Average *scale discounts* in per-meal bid prices placed during 1999 - 2005 (Olivares et al., 2011). Dashed lines indicate 95% confidence interval of the estimates. The discount is measured by the decrease in the per-meal bid price when individual units are combined into a multi-unit package in the corresponding volume. For example, increasing package size to 20 million meals (combining about 8 units) generates discounts of around 6% of average bid price. All the bid prices are normalized to 1999 values using consumer price index.

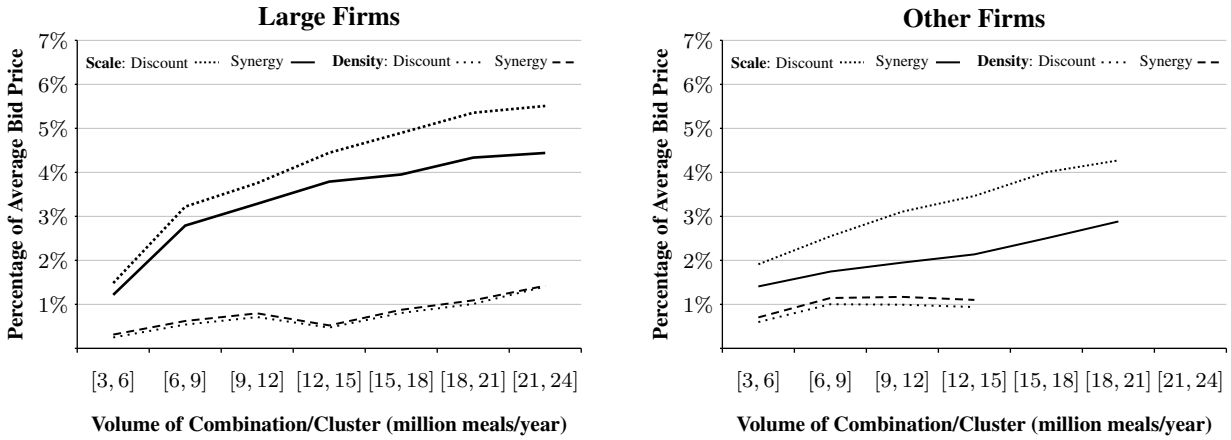


Figure 2 – Estimates of the discount curves (equation (8)) and the synergy curves (equation (10)) for each group of firms for 2003 auction. Scale discounts and synergies are drawn against combination size and density discounts and synergies are drawn against cluster size. The estimation is done with 5101 observations.

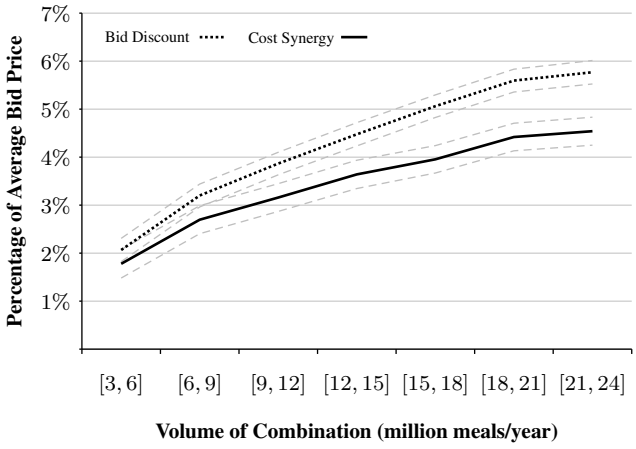


Figure 3 – Estimates of the discount curves (equation (8)) and the synergy curves (equation (10)) without density terms for 2003 auction. 95% confidence intervals are shown in dashed lines. The number of observations is 5101.

Online Appendix

A Winner Determination Problem Formulation

In this section, we will provide the details of the integer programing (IP) formulation of the winner determination problem (WDP). This IP is formulated and solved in the course of simulation described in Section 4.2 to estimate each firm's markup vector. We begin by introducing notation that is not defined in the main body of the paper and we then formulate the IP.

Index Sets. We let R denote the set of geographical regions indexed by r (recall that each geographical region contains several TUs). We let $\mathcal{A}_f$ be the set of packages on which firm f places bids. They are to distinguish from $\mathcal{A}$ in case of missing bids (unobserved bids) by firm f. $\mathcal{A}_{rf} \subseteq \mathcal{A}_f$ represents the set of packages in $\mathcal{A}_f$ that contain at least one TU in region r. Finally, we let $|a|$ denote the number of TUs in package a, and we let A_f and A_{rf} denote the number of packages in the sets $\mathcal{A}_f$ and $\mathcal{A}_{rf}$, respectively.

Decision Variables. We let x_{af} be the firm-package allocation decision variable for package a and firm f. This variable takes the value of 1, if firm f wins package a, and 0 otherwise. These variables determine the final allocation. The variable y_{rf} is a regional allocation variable for region r and firm f, taking the value of 1 if firm f wins a package that contains at least one TU in region r, and 0 otherwise. They are used to count the number of firms serving in each geographical region for the local constraints. The decision variable z_f relates to the winning status of firm f. It is equal to 1 if firm f wins a package and 0 otherwise. They count the number of winning firms to be used in the global competition constraint.

Constraints and Their Parameters. As described in Section 3, we have five types of allocative constraints in the auction. We also have an additional constraint imposed in our structural model, namely, that each firm can win *at most one* package. We label those constraints as follows: (A) *Cover all TUs* ensures that all the TUs be contracted. (B) *At most one package* constraint imposes that firms can win at most one package. (C) *Maximum number of TUs* bounds the number of TUs that each firm can win. We let MXU_f denote the maximum number of TUs that firm f can win. (D) *Global Market Share Constraints* limits the total volume of standing contracts of each firm in terms of the number of meals served. We let MXM_f denote the total number of meals that firm f can win in the auction being considered. (E) *Local constraints* bound the minimum and maximum number of firms serving in each region. We use MNF_r and MXF_r to denote these bounds for region r. (F) *Global competition constraint* sets the minimum number of firms being contracted in the auction being considered. We let MNF_g denote this minimum number.

Notice that constraints (C) and (D) are firm-wise limits, and for each firm any bids placed on packages that exceed the firm's limits can never win. Therefore, we eliminate such bids *a priori* from $\mathcal{A}_f$ for each firm $f \in F$. That is for any given firm f and for all $a \in \mathcal{A}_f$, we have $|a| \leq MXU_f$ and $v_a \leq MXM_f$. Then, constraints (C) and (D) will be automatically satisfied as long as firms win at most one package imposed by (B). Hence, we omit (C) and (D) in our IP formulation. Recall that the objective is to minimize the total procurement cost.

Now we present the IP formulation of the WDP. The constraints that are not labeled impose the correct values for the auxiliary variables y_{rf} and z_f, and the integrality constraints for all decision variables.

$$
\begin{aligned}
&\textbf{minimize} && \sum_{f\in F}\sum_{a\in\mathcal{A}_f} b_{af}x_{af} \\
&\textbf{subject to} \quad (A) && \sum_{f\in F}\sum_{a\in\mathcal{A}_f:i\in a} x_{af} \geq 1, \quad \forall i \in U \\
&\qquad\quad (B) && \sum_{a\in\mathcal{A}_f} x_{af} \leq 1, \quad \forall f \in F \\
&\qquad\quad (E) && MNF_r \leq \sum_{f\in F} y_{rf} \leq MXF_r, \quad \forall r \in R \\
&\qquad && \frac{1}{A_{rf}} \sum_{a\in\mathcal{A}_{rf}} x_{af} \leq y_{rf} \leq \sum_{a\in\mathcal{A}_{rf}} x_{af}, \quad \forall r \in R,\ \forall f \in F \\
&\qquad\quad (F) && \sum_{f\in F} z_f \geq MNF_g, \\
&\qquad && \frac{1}{A_f} \sum_{a\in\mathcal{A}_f} x_{af} \leq z_f \leq \sum_{a\in\mathcal{A}_f} x_{af}, \quad \forall f \in F \\
&\qquad && x_{sf}, y_{rf}, z_f \in \{0,1\}.
\end{aligned}
$$

B Proofs

B.1 Notation

We begin by defining notation that is frequently used in this section. First we consider a focal bidder f, whose observed bid vector is b. All of the analysis is focused on this bidder, and as before we omit the index f whenever it is clear from the context. Recall that from the perspective of this focal bidder, competitors' bid prices are random. All such random quantities are defined over a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Note that $\mathbf{P}$ measures the probability of each of the events characterized by the final allocation of units to bidders in the CA. Hence, it defines the vector of winning probabilities $G(\cdot)$. In addition, we define $\Omega^* \subseteq \Omega$ to be the sample space where ties never happen in the winner determination problem. By Assumption 4, the distribution of competitors' bids is absolutely continuous, and hence, we can find such a sample space so that $\mathbf{P}(\Omega^*) = 1$. In words, this means that the winner determination problem has a *unique* solution for any realization $\omega \in \Omega^*$. Therefore, in our analysis we do not consider any issues related to tie-breaking in the final allocation.

We let b' be the vector of competitors' bid prices. That is, given a realization of $\omega \in \Omega^*$, $b'(\omega) = \{b'_{f'}(\omega)\}_{f'\neq f}$, where $b'_{f'}(\omega)$ is a vector of bids for competing firm f'. Furthermore, we let $x = \{x_{af}\}_{a\in\mathcal{A},f\in F}$ be a $A \times |F|$ dimensional vector such that x_{af} takes 1 if bidder f wins package a and 0 otherwise. A vector x uniquely determines an allocation outcome. We denote by X, the set of all feasible allocation outcomes that satisfy all the allocative constraints in the CA including the one that each bidder can win at most one package (see Assumption 1). All the proofs in this section are valid under any other allocative constraints in the CA as long as they do not depend on bid prices (so the constraints described in Section 3.2 are all valid). In addition, we let $X_a \subset X$ be the set of allocations such that bidder f wins package a. We adopt

the null package indexed by 0, and accordingly, we use X_0 to denote the set of allocations in which bidder f wins no package, and let $G_0(b)$ be the probability that bidder f wins no package given her bid vector b. Note that because bidders can win at most one package, X_a and X_s are disjoint for any packages $a \neq s$ in $\mathcal{A}_0 = \mathcal{A} \cup \{0\}$, and we have $\bigcup_{a \in \mathcal{A}_0} X_a = X$.

Without loss of generality, we assume x is ordered in a way such that the vector of bidder f's allocation decisions, denoted by x_f, is followed by the vector of competitors' allocation decisions, denoted by x', so that $x = (x_f, x')$. Additionally, we define a cost function: $p_a(\omega) := \min_{x \in X_a}(b, b'(\omega))^T x$, for each $a \in \mathcal{A}_0$. This is the minimum total procurement cost out of all the allocations such that bidder f wins package a given a realization $\omega \in \Omega^*$. It is important to note that because of the constraint that each bidder can win at most one package, $p_a(\omega)$ only depends on the value of b_a among bidder f's bids for all $a \in \mathcal{A}$.

Finally, for notational simplicity, we use $G_{a,s}(b)$ to denote the partial derivative of the winning probability $G_a(b)$ with respect to the bid price b_s. In addition, when dealing with a characteristic-based markup model, we let $\mathcal{A}_i \subseteq \mathcal{A}$ to denote the set of packages associated with the i^{th} markup variable, and also let $G_{a,\theta_i}(b)$ to denote the partial derivative of the winning probability $G_a(b)$ with respect to the markup variable θ_i.

B.2 Proofs

We will use some lemmas for the proofs of the main results. The following lemma is useful for the proof of Proposition 1. The result follows by applying the fundamental theorem of calculus and its proof is omitted.

Lemma B.1. *Define a function $F : \Re^n \mapsto \Re$ such that:*

$$F(y) = \int_{D(y)} f(x)dx,$$

where $f : \Re^m \mapsto \Re$ is continuous and integrable in $\Re^m$. Assume that the domain of integration $D(y)$ is a polyhedron formed by a given matrix $A \in \Re^{k \times m}$ and a vector function $b(y) \in \Re^k$ with $k \in \mathbb{N}$ such that $D(y) := \{x \in \Re^m : Ax \leq b(y)\}$. If $b(y)$ is differentiable with respect to y, then F is continuous and differentiable everywhere in $\Re^n$.

Proof of Proposition 1. To prove the differentiability of the winning probability vector $G(b)$ with respect to b, we begin by considering an arbitrary package $a \in \mathcal{A}$ and look at the winning probability that bidder f wins package a, $G_a(b)$. Notice that bidder f wins package a if one of the allocations in X_a achieves the minimum procurement cost among all possible allocations in X. We let $K := |X_a|$, the number of distinct allocations in X_a, and index them by $k = 1, 2, \cdots, K$. Specifically, we look at the event that bidder f wins package a as a result of allocation $x_k \in X_a$. Accordingly, we let $G_a(b; x_k)$ denote the probability that $x_k \in X_a$ becomes the final allocation (hence the minimizer of the total procurement cost). Because the probability of ties is zero, the winning probability $G_a(b)$ can be expressed as $G_a(b) = \sum_{k=1}^{K} G_a(b; x_k)$. Therefore it suffices to show that $G_a(b; x_k)$ is continuous and differentiable for any given allocation x_k.

Now given an arbitrary allocation $x_k \in X_a$, we show the differentiability of $G_a(b; x_k)$ using Lemma B.1. By letting $f(b')$ denote the joint probability density function of competitors' bids b', the winning probability $G_a(b; x_k)$ can be written as: $G_a(b; x_k) = \int_{D_k(b)} f(b')db'$, where $D_k(b)$ is the set of b''s for

which x_k is the optimal allocation given b. Observe that $D_k(b)$ can be expressed by a set of inequalities as follows:

$$x_k^T(b,b') \leq y^T(b,b'), \ \ \forall y \in X \quad (\Rightarrow) \quad (x'_k - y')^T b' \leq (y_f - x_{kf})^T b, \ \ \forall y \in X.$$

The inequalities ensure that, given the placed bids (b, b'), the total procurement cost incurred by allocation x_k is cheaper than that of any other feasible allocations if we do not consider ties. Therefore, we get: $D_k(b) = \{b' \in \Re^{A\times(|F|-1)} : (x'_k - y')^T b' \leq (y_f - x_{kf})^T b, \forall y \in X\}$. If we let $J = |X|$ and index the feasible allocations by j, then $D_k(b)$ is a polyhedron in $\Re^{A\times(|F|-1)}$ defined by $Mb' \leq h(b)$, where the j^{th} row of M is $(x'_k - y'_j)^T$ and the j^{th} element of vector $h(b)$ is $(y_f - x_{kf})^T b$, for $j = 1, \cdots, J$.

By Assumption 4, the density $H(\cdot|Z_{f'})$ is continuous everywhere and independent across bidders, and hence, the joint density $f(b')$ is continuous on $\Re^{A\times(F-1)}$. The integrability of $f(b')$ is readily obtained as it is a probability density function. Finally, the function $h(b)$ is a linear function of b, and hence differentiable with respect to b. Therefore, by Lemma B.1, $G_a(b; x_k)$ is continuous and differentiable with respect to the bid vector b. Since the choice of package $a \in \mathcal{A}$ and allocation $x_k \in X_a$ was arbitrary, the proof is complete. ■

It is useful to examine some of the properties of the Jacobian matrixes $\mathcal{D}_b G(b)$ and $\mathcal{D}_\theta W^T G(b)$ for the proof of Proposition 2 and Theorem 1. The following lemma investigates those properties.

Lemma B.2. *For any given bidder and her bid vector b, we have the following properties for the winning probability vector $G(b)$.*

1. *The Jacobian matrix $\mathcal{D}_b G(b)$ is symmetric.*
2. *For any package $a \in \mathcal{A}$, we have i) $G_{a,a}(b) \leq 0$; ii) $G_{s,a}(b) \geq 0$ for all $s \in \mathcal{A} \setminus \{a\}$; and iii) $\sum_{s\in\mathcal{A}} G_{s,a}(b) \leq 0$.*
3. *Consider a group markup model specified by a package-characteristic matrix W whose elements are all non-negative. Let the markup vector θ and $D := \mathcal{D}_\theta W^T G(b)$. Then $D_{ij} \geq 0$ for any $i \neq j$.*

Proof. (Proof of Part 1). We fix two arbitrary but distinct packages a and s, and we first show that $G_{s,a}(b) \leq G_{a,s}(b)$. We then establish the reversed inequality by exchanging the two packages and using a symmetric argument. The arbitrary choice of the two packages a and s then provides the completion of the proof.

Accordingly, take any two distinct packages $a, s \in \mathcal{A}$ and an arbitrary scalar $\epsilon > 0$. We begin by defining the following events:

$$\begin{aligned}
\Omega_a &:= \{\omega \in \Omega^* : p_a(\omega) = \min_{t\in\mathcal{A}_0} p_t(\omega)\}, \\
\Omega_{a,s} &:= \{\omega \in \Omega_a : p_s(\omega) = \min_{t\in\mathcal{A}_0\setminus\{a\}} p_t(\omega)\}, \\
\Omega^\epsilon_{a,s} &:= \{\omega \in \Omega_a : p_s(\omega) < p_a(\omega) + \epsilon\}.
\end{aligned}$$

By definition, Ω_a denotes the event where bidder f wins package a, and $\Omega_{a,s} \subset \Omega_a$ denotes the events where the minimum allocation without bidder f winning package a is the one with her winning package s.

Also, $\Omega^{\epsilon}_{a,s} \subset \Omega_a$ is the event where the minimum allocation with bidder f winning package s is less than ϵ above from the optimal value, $p_a(\omega)$. Finally we let $\Omega_s \subset \Omega^*$ to be the event where bidder f wins package s. Note that Ω_a and Ω_s are disjoint.

We use the following random variables: $Y_s^{a\pm\epsilon}(\omega) := \mathbf{1}[\ p_s(\omega) = \min(\min_{t\in\mathcal{A}_0\setminus\{a\}} p_t(\omega), p_a(\omega)\pm\epsilon)\ ]$. The random variables $Y_s^{a\pm\epsilon}(\omega)$ indicate bidder f's winning of package s when her bid b_a changes by $+\epsilon$ and $-\epsilon$, respectively. Similarly, we define $Y_s^0(\omega) := \mathbf{1}[\ p_s(\omega) = \min_{t\in\mathcal{A}_0} p_t(\omega)]$, that is, the indicator that the bidder wins package s given her bid price b at the realization of ω. Now we divide the event set Ω^* into the following four disjoint subsets and examine the values of the random variables $Y_s^{a+\epsilon}(\omega)$ and $Y_s^0(\omega)$.

1. $\forall\omega \in \Omega^* \setminus (\Omega_a \cup \Omega_s)$: Bidder f is winning neither a nor s, so $Y_s^0(\omega) = 0$. Moreover, increasing her bid b_a by ϵ will not let her win s, hence, $Y_s^{a+\epsilon}(\omega) = 0$.

2. $\forall\omega \in \Omega_s$: Bidder f is winning package s and increasing her bid on non-winning package a will not change her winning s. Thus, $Y_s^0(\omega) = Y_s^{a+\epsilon}(\omega) = 1$.

3. $\forall\omega \in \Omega_{a,s} \cap \Omega^{\epsilon}_{a,s}$: Bidder f is winning package a, so $Y_s^0(\omega) = 0$. Since $\omega \in \Omega^{\epsilon}_{a,s}$, after increasing b_a by ϵ, the value of the current optimal allocation $p_a(\omega) + \epsilon$ becomes larger than $p_s(\omega)$. But then, $\omega \in \Omega_{a,s}$ implies $p_s(\omega)$ becomes the lowest cost allocation after such a perturbation. Hence, $Y_s^{a+\epsilon}(\omega) = 1$.

4. $\forall\omega \in \Omega_a \setminus (\Omega_{a,s} \cap \Omega^{\epsilon}_{a,s})$: Bidder f is winning package a, so $Y_s^0(\omega) = 0$. If $\omega \notin \Omega_{a,s}$, after increasing b_a by ϵ, $p_s(\omega)$ is not the lowest cost allocation. If $\omega \notin \Omega^{\epsilon}_{a,s}$, $p_s(\omega)$ is still larger than the value of the current allocation, $p_a(\omega) + \epsilon$, even after the perturbation. Hence, $Y_s^{a+\epsilon}(\omega) = 0$.

In words, $(\Omega_{a,s} \cap \Omega^{\epsilon}_{a,s})$ is the only event in which bidder f's winning status of package s changes by an ϵ increase in her bid b_a. Therefore, we obtain:

$$\frac{G_s(b + \epsilon e_a) - G_s(b)}{\epsilon} = \frac{1}{\epsilon}\mathbf{E}[Y_s^{a+\epsilon} - Y_s^0] = \frac{1}{\epsilon}\mathbf{P}(\Omega_{a,s} \cap \Omega^{\epsilon}_{a,s}), \tag{B.1}$$

where e_a is the a^{th} canonical vector whose a^{th} component is the only non-zero element and is equal to one.

Now we look at the effect of decreasing b_s by ϵ to the winning of package a. Similarly, we divide the event set Ω^* into the following three disjoint subsets and examine the values of random variables $Y_a^{s-\epsilon}(\omega)$ and $Y_a^0(\omega)$.

1. $\forall\omega \in \Omega^* \setminus (\Omega_a)$: Since bidder f is not winning package a, $Y_a^0(\omega) = 0$. Moreover, decreasing her bid b_s by ϵ will never let her win package a, hence, $Y_a^{s-\epsilon}(\omega) = 0$.

2. $\forall\omega \in \Omega^{\epsilon}_{a,s}$: Bidder f is winning package a, so $Y_a^0(\omega) = 1$. Since $\omega \in \Omega^{\epsilon}_{a,s}$, after decreasing b_s by ϵ, $p_s(\omega) - \epsilon$ has a lower cost than the current optimal value, $p_a(\omega)$, so bidder f will win package s instead of a. Hence, $Y_a^{s-\epsilon}(\omega) = 0$.

3. $\forall\omega \in (\Omega_a \setminus \Omega^{\epsilon}_{a,s})$: Bidder f is winning package a, so $Y_a^0(\omega) = 1$. Since $\omega \notin \Omega^{\epsilon}_{a,s}$, decreasing b_s by ϵ cannot make the value $p_s(\omega) - \epsilon$ better than the current optimal value, $p_a(\omega)$. Hence, the previous optimal allocation will remain optimal and $Y_a^{s-\epsilon}(\omega) = 1$.

This time, $\Omega^{\epsilon}_{a,s}$ is the only case that bidder f's winning status of package a is affected by an ϵ decrease in her bid b_s. Therefore:

$$\frac{G_a(b) - G_a(b - \epsilon e_s)}{\epsilon} = \frac{1}{\epsilon}\mathbf{E}[Y^0_a - Y^{s-\epsilon}_a] = \frac{1}{\epsilon}\mathbf{P}(\Omega^{\epsilon}_{a,s}). \tag{B.2}$$

Since $(\Omega_{a,s} \cap \Omega^{\epsilon}_{a,s}) \subseteq \Omega^{\epsilon}_{a,s}$, from (B.1) and (B.2) we get the following inequality:

$$\frac{G_s(b + \epsilon e_a) - G_s(b)}{\epsilon} = \frac{1}{\epsilon}\mathbf{P}(\Omega_{a,s} \cap \Omega^{\epsilon}_{a,s}) \leq \frac{1}{\epsilon}\mathbf{P}(\Omega^{\epsilon}_{a,s}) = \frac{G_a(b) - G_a(b - \epsilon e_s)}{\epsilon}.$$

Recall that ϵ is an arbitrary positive scalar and Proposition 1 ensures the differentiability of $G(b)$ with respect to b. Thus, by letting ϵ vanish, we get $G_{s,a} \leq G_{a,s}$.

In the previous argument, the only condition for the packages a and s is that they are distinct. Hence, a symmetric argument also holds true and we get $G_{s,a} \geq G_{a,s}$, and therefore we get $G_{s,a} = G_{a,s}$. The arbitrary choice of a and s then let us conclude $G_{a,s} = G_{s,a}$, for any two distinct packages $a, s \in \mathcal{A}$. This completes the proof of part 1.

(Proof of Part 2). To show $G_{a,a}(b) \leq 0$, fix a realization of $\omega \in \Omega^*$ and consider a perturbation of increasing bidder f's bid price b_a by $\epsilon > 0$. If she currently wins package a, she may or may not win package a after the perturbation. However, if she currently does not win package a, i.e., $p_a(\omega)$ is not the lowest cost allocation, she cannot win package a after the perturbation since $p_a(\omega) + \epsilon$ remains being larger than the current optimal value. Since these are true for any $\omega \in \Omega^*$, increasing bid price b_a will never increase her chances of winning package a. Hence we get $G_a(b + \epsilon e_a) \leq G_a(b)$, for all $\epsilon > 0$. Then the differentiability of $G(b)$, shown in Proposition 1, implies $G_{a,a}(b) \leq 0$.

Similarly, for the proof of $G_{s,a}(b) \geq 0$ for any $s \in \mathcal{A} \setminus \{a\}$, consider a perturbation of decreasing b_a by an arbitrary $\epsilon > 0$. Given a realization of $\omega \in \Omega^*$, if she currently wins package s (possibly the null package), she can either win package a instead of s or still win package s after the perturbation. However, if she currently wins package a, she will win package a for sure after the perturbation. Therefore, decreasing her bid b_a only possibly decrease her chances of winning package s, and we get $G_s(b) \geq G_s(b - \epsilon e_a)$, for all $\epsilon > 0$. Again by the differentiability of $G(b)$, we get $G_{s,a}(b) \geq 0$.

Finally, since $\sum_{s \in \mathcal{A}} G_s(b) = 1 - G_0(b)$, so we get $\sum_{s \in \mathcal{A}} G_{s,a}(b) = -G_{0,a}(b) \leq 0$, where the last inequality follows because $G_{0,a}(b) \geq 0$ by a similar argument than above. This completes the proof of part 2.

(Proof of Part 3). Note that by Assumption 5, $b = W\theta + c$ and by the chain rule, we have $D := \mathcal{D}_\theta W^T G(b) = W^T \mathcal{D}_b G(b) W$. Then for any $i \neq j$ we get:

$$D_{ij} = \sum_{a,s \in \mathcal{A}} W_{ai} W_{sj} G_{a,s}(b) = \sum_{a \in \mathcal{A}_i, s \in \mathcal{A}_j} W_{ai} W_{sj} G_{a,s}(b),$$

where the second equality comes from the fact that $W_{ai} = 0$ if $a \notin \mathcal{A}_i$ by its definition. In addition, recall that in the group markup model, $\mathcal{A}_i$ and $\mathcal{A}_j$ are disjoint if $i \neq j$. Therefore by part 2 of this lemma shown above, $G_{a,s}(b) \geq 0$ for all $a \in \mathcal{A}_i$ and $s \in \mathcal{A}_j$. The non-negativity of the elements in W then ensures that $D_{ij} \geq 0$ for all $i \neq j$, which completes the proof of part 3. ■

Proof of Proposition 2. In the full-dimensional markup model, we have $b_a = c_a + \theta_a$ for $a = 1, \ldots, K$, and the first-order conditions, (3) yields:

$$[\mathcal{D}_b G(b)]^T \theta = -G(b), \text{ where } \theta = [\theta_1, ..., \theta_K]^T. \tag{B.3}$$

Similarly, for the common markup specification, we have $b_a = c_a + \theta_u$ for all $a = 1, \ldots, K$. Note that the package-characteristic matrix $W \in \Re^K$ is then $W = [1, 1, ..., 1]^T$. By letting $\alpha := [\alpha_1, ..., \alpha_K]^T$ where $\alpha_a := G_{a,\theta_u}(b)$, we have $\mathcal{D}_{\theta_u} W^T G(b) = W^T \mathcal{D}_{\theta_u} G(b) = W^T \alpha$. Then the first-order condition of this characteristic-based markup model, (5) now becomes:

$$[\mathcal{D}_{\theta_u} W^T G(b)]^T \theta_u = -W^T G(b) \quad (\Rightarrow) \quad \alpha^T W \theta_u = -W^T G(b). \tag{B.4}$$

Observe that by definition, $\frac{\partial b_s}{\partial \theta_u} = 1$ for all $s = 1, 2, ..., K$. Therefore by the chain rule, we get:

$$\alpha_a = G_{a,\theta_u}(b) = \sum_{s=1}^{K} G_{a,s}(b) \quad (\Rightarrow) \quad W^T [\mathcal{D}_b G(b)]^T = \alpha^T.$$

Using this, left-multiplying by W^T on both sides of equation (B.3) and then equating the right-hand sides of equations (B.3) and (B.4) yields:

$$\sum_{a=1}^{K} \alpha_a \theta_a = \left(\sum_{a=1}^{K} \alpha_a \right) \theta_u \quad (\Rightarrow) \quad \theta_u = \frac{1}{\sum_{a=1}^{K} \alpha_a} \sum_{a=1}^{K} \alpha_a \theta_a.$$

Finally, by symmetry of the Jacobian matrix $\mathcal{D}_b G(b)$, shown in part 1 of Lemma B.2, we have $\alpha_a = \sum_{s=1}^{K} G_{a,s} = \sum_{s=1}^{K} G_{s,a}$. Then part 2 of the same lemma implies $\alpha_a \leq 0$ for all $a = 1, ..., K$. We end the proof by showing that at least one $\alpha_a < 0$. Assume for the purpose of contradiction that α_a's are all zero. This implies that the sum of all the column vectors in the Jacobian matrix $\mathcal{D}_b G(b)$ is a zero vector and therefore they are not linearly independent. However, since all the bids have strictly positive winning probabilities, the Jacobian matrix $\mathcal{D}_b G(b)$ is invertible as will be shown in Theorem 1, hence a contradiction. Therefore, we conclude that at least one α_a is strictly negative, and this completes the proof. ■

Proof of Proposition 3. The first-order conditions of the full-dimensional model, (3) gives:

$$\begin{bmatrix} G_{1,1}(b) & G_{2,1}(b) & G_{12,1}(b) \\ G_{1,2}(b) & G_{2,2}(b) & G_{12,2}(b) \\ G_{1,12}(b) & G_{2,12}(b) & G_{12,12}(b) \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_{12} \end{bmatrix} = - \begin{bmatrix} G_1(b) \\ G_2(b) \\ G_{12}(b) \end{bmatrix} \tag{B.5}$$

Now consider the case where we use common markup θ_u for single unit bids and markup θ_v for the package of units 1 and 2, so that the package-characteristic matrix W is formed as follows:

$$W = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{matrix} \rightarrow \text{ Unit 1: apply unit markup } \theta_u, \\ \rightarrow \text{ Unit 2: apply unit markup } \theta_u, \\ \rightarrow \text{ Package 12: apply package markup } \theta_v. \end{matrix}$$

Note that by the chain rule, $G_{a,\theta_u}(b) = G_{a,1}(b) + G_{a,2}(b)$, for $a = 1, 2, 12$. Using this, the first-order condition of the characteristic-based model, (5) yields:

$$\begin{bmatrix} G_{1,1}(b) + G_{1,2}(b) + G_{2,1}(b) + G_{2,2}(b) & G_{12,1}(b) + G_{12,2}(b) \\ G_{1,12}(b) + G_{2,12}(b) & G_{12,12}(b) \end{bmatrix} \begin{bmatrix} \theta_u \\ \theta_v \end{bmatrix} = - \begin{bmatrix} G_1(b) + G_2(b) \\ G_{12}(b) \end{bmatrix} \tag{B.6}$$

By left-multiplying by W^T on both sides of (B.5) and then equating the right-hand sides of equations (B.5) and (B.6), we get the desired relationship of the two markup vectors. Note that by Theorem 1, the Jacobian matrix in (B.6) is invertible and therefore $det \neq 0$. This completes the proof. ■

Proof of Lemma 1. Fix a package $a \in \mathcal{A}$. Note that by the chain rule and Assumption 5, we have $G_{a,\theta_i}(b) = \sum_{s\in\mathcal{A}} \frac{\partial b_s}{\partial \theta_i} G_{a,s}(b) = \sum_{s\in\mathcal{A}} W_{si} G_{a,s}(b)$. Therefore, it suffices to show that $G_{a,s}(b) = 0$ for all $s \in \mathcal{A}$.

First, we let $\underline{p}(\omega) := \min_{t\in\mathcal{A}_0} p_t(\omega)$, the minimum procurement cost given $\omega \in \Omega^*$. Note that $G_a(b) = 0$ implies $p_a(\omega) > \underline{p}(\omega)$ in a set of $\Omega_a \subseteq \Omega^*$, such that $\mathbf{P}(\Omega_a) = 1$. Also, we let $e_a \in \Re^A$ be the a^{th} canonical vector, whose a^{th} component is equal to one while all others are equal to zero.

We now show that $G_{a,s}(b) = 0$ for all $s \in \mathcal{A} \setminus \{a\}$. First, take any package $s \neq a$ and consider a perturbation of decreasing b_s by $\epsilon > 0$. Recall that bidder f can win at most one package and therefore $p_a(\omega)$ does not depend on the value of b_s. Therefore, decreasing b_s will not change the value of $p_a(\omega)$. However, depending on whether b_s is part of the current optimal allocation or not, the value of the current optimal allocation may decrease by ϵ or stay the same ($\underline{p}(\omega)$) after the perturbation. Thus, after such a perturbation the value of the current allocation will still have a lower cost than $p_a(\omega)$. This implies that bidder f remains not winning package a for all $\omega \in \Omega_a$. Hence, we obtain $G_a(b) - G_a(b - \epsilon e_s) = 0$ for all $\epsilon > 0$. Then the differentiability of $G(b)$ established in Proposition 1 implies $G_{a,s}(b) = 0$.

Similarly, to show that $G_{a,a}(b) = 0$, consider a perturbation of increasing b_a by $\epsilon > 0$. Then again for all $\omega \in \Omega_a$, after such a perturbation, $p_a(\omega)$ only increases to be $p_a(\omega) + \epsilon$, and remains being larger than the optimal value $\underline{p}(\omega)$. Hence bidder f can never win package a after the perturbation, which implies $G_a(b + \epsilon e_a) - G_a(b) = 0$ for all $\epsilon > 0$. Again by Proposition 1, we obtain $G_{a,a}(b) = 0$.

By combining these results, we finally get $G_{a,\theta_i}(b) = \sum_{s\in\mathcal{A}} W_{si} G_{a,s}(b) = 0$, which completes the proof. ■

The following Lemma provides invertibility conditions of a matrix, which is used to prove Theorem 1.

Lemma B.3 (Theorem 6.1.10 in Horn and Johnson (1985))**.** *A matrix $D \in \Re^{n\times n}$ is said to be* strictly diagonally dominant*, if it satisfies:*

$$|D_{ii}| > \sum_{j\neq i} |D_{ij}|, \quad \forall i = 1, 2, \ldots, n.$$

If D is strictly diagonally dominant, then D is invertible.

Proof of Theorem 1: (Necessity). We first show that if the Jacobian matrix $\mathcal{D}_\theta W^T G(b)$ is invertible it must be that every markup variable has at least one relevant bid associated with it. For this, assume there exists a markup variable, say θ_i, whose associated bids are all irrelevant. Now note that $[\mathcal{D}_\theta W^T G(b)]_{ij} =$

$\sum_{a\in\mathcal{A}_i} W_{ai}G_{a,\theta_j}(b)$. But then Lemma 1 implies that $G_{a,\theta_j}(b) = 0,\ \forall a \in \mathcal{A}_i$, leading to $[\mathcal{D}_\theta W^T G(b)]_{ij} = 0$. Since this is true for any $j = 1, 2, ..., d$, the i^{th} row of Jacobian matrix $\mathcal{D}_\theta W^T G(b)$ will be a zero vector. Having a row of zeros implies that the matrix is not invertible. This completes the proof of necessity. ■

Proof of Theorem 1: (Sufficiency). We now show that if every markup variable has at least one relevant bid associated with it and the additional conditions in the statement of the theorem hold, then the Jacobian matrix $\mathcal{D}_\theta W^T G(b)$ evaluated at the observed bid vector b is invertible, and therefore the markup vector θ is uniquely determined by equation (6). For notational simplicity, we let $D := \mathcal{D}_\theta W^T G(b)$.

First, recall that in a group markup specification, for any package a, there is only one markup variable that is associated with it, say markup variable θ_i. Then the profit that bidder f makes from winning package a is $W_{ai}\theta_i$. By assumption, $W_{ai}\theta_i \geq 0$ and $W_{ai} \geq 0$, for all packages a. Therefore, $\theta_i \geq 0$, for all i. We now proceed to show that θ_i is indeed strictly positive for all $i = 1, 2, ..., d$. By assumption, θ satisfies equation (5): $D^T\theta = -W^T G(b)$. For the purpose of contradiction, we fix i and assume that θ_i is zero. We examine the i^{th} equation in (5):

$$D_{ii}\theta_i + \sum_{j\neq i} D_{ji}\theta_j = -W_i^T G(b). \tag{B.7}$$

The first term on the left-hand side is zero by assumption. The second term is non-negative since we know that (i) $\theta_j \geq 0,\ \forall j$; and (ii) $D_{ji} \geq 0$ by part 3 of Lemma B.2. However, the right-hand side is strictly negative because there is at least one relevant bid, say b_a, that is associated with markup variable θ_i, so that $W_i^T G(b) \geq W_{ai}G_a(b) > 0$. Therefore it is impossible for θ to satisfy equation (5), which contradicts our assumption. Hence, $\theta_i > 0$, for all i.

Now, we construct a diagonal matrix Θ so that $\Theta_{ii} = \theta_i$ for all $i = 1, 2, ..., d$. Because $\theta_i > 0,\ \forall i$, it is clear that Θ is invertible. We now show that equation (5) implies that the matrix $D^T\Theta$ is strictly diagonally dominant, and therefore invertible by Lemma B.3. To see this, take any $i \in \{1, 2, ..., d\}$, and consider the i^{th} equation in (5) (see (B.7)), for which we know that its right-hand side is strictly negative. Therefore, using $[D^T\Theta]_{ij} = D_{ji}\Theta_{jj} = D_{ji}\theta_j$, we reach the following inequality:

$$[D^T\Theta]_{ii} + \sum_{j\neq i}[D^T\Theta]_{ij} = -W_i^T G(b) < 0 \quad (\Rightarrow) \quad \sum_{j\neq i}[D^T\Theta]_{ij} < -[D^T\Theta]_{ii}.$$

Recall that when $i \neq j$, we have $[D^T\Theta]_{ij} = D_{ji}\theta_j \geq 0$, and this implies $\sum_{j\neq i} |[D^T\Theta]_{ij}| < |[D^T\Theta]_{ii}|$. Since this is true for any $i = 1, 2, ..., d$, we conclude that $D^T\Theta$ is strictly diagonally dominant and hence invertible by Lemma B.3. Since Θ is also invertible, the invertibility of D follows with $D^{-1} = (\Theta^T D)^{-1}\Theta^T$. The proof for sufficiency is now complete. ■

C VCG Payment Rule and a Core Outcome.

C.1 VCG Payment Rule

First, we describe the payment rules of the VCG mechanism, which we then use to calculate total payments under VCG. Let $V(F)$ denote the value of the minimum-cost allocation that satisfies all constraints based on the reported bids of all firms in set F. Because VCG is truthful, these bids correspond to actual costs. In

addition, let $F^* \subseteq F$ be the set of firms who are awarded contracts in the VCG allocation and let $b_{a(f),f}$ be the bid price reported by firm $f \in F^*$ for her winning package $a(f)$ (in this notation $b_{a(f),f}$ represents the total value for the entire package, not the per-meal value). The VCG payment to winner $f \in F^*$, denoted by P_f, is computed as follows:

$$P_f = V(F_{-f}) - \sum_{f' \in F^*_{-f}} b_{a(f'),f'},$$

where $F_{-f} = F \backslash \{f\}$ and $F^*_{-f} = F^* \backslash \{f\}$. The first term is the total value of reported bids in the optimal allocation that considers all bids except those from winning firm f. The second term is the total value of reported bids in the current VCG allocation (that includes firm f), except for the reported value of firm f's winning package. Hence, the payment to a winner is essentially the cost of providing the units she wins in the lowest cost allocation without her. Loosing bidders do not receive payments. The total procurement cost for the auctioneer under VCG is then obtained by summing up all such individual payments to winning firms.

C.2 Finding a Core Outcome Close to VCG

Now we turn our attention to the concept of a core outcome in a CA. Specifically we are interested in checking whether the VCG outcome lies in the core or whether it is *close* to it. We start by providing some useful definitions. We closely follow Day and Raghavan (2007); Day and Milgrom (2008) also provide a useful description of this material. First, we call the final allocation and the payments to bidders in a CA an *outcome*. Given an outcome, Γ, we call the set of winning bidders a *coalition*, C_Γ. An outcome Γ is said to be *blocked* if there exists an alternative outcome $\overline{\Gamma}$ that generates strictly lower total procurement cost to the auctioneer and for which every bidder in $C_{\overline{\Gamma}}$ weakly prefers $\overline{\Gamma}$ to Γ. An efficient outcome Γ that is not blocked, is called a *core outcome*. Note that if an outcome is not in the core, there is a group of bidders that have incentives to deviate from it and offer a better deal to the auctioneer.

In addition, a core outcome Γ is called *bidder-Pareto optimal* if there is no other core outcome weakly preferred by every bidder in C_Γ. Day and Raghavan (2007) and Day and Milgrom (2008) propose auctions that find efficient, core, bidder-Pareto optimal outcomes. An attractive property of efficient core-selecting auctions that are also bidder-Pareto optimal is that they minimize the incentives to unilaterally misreport true costs among all core-selecting auctions. In this sense, these auctions have outcomes that are *closest* to VCG among all core outcomes. We use the algorithm proposed by Day and Raghavan (2007) to find a core outcome that is closest to VCG.[15]

D Counterfactual Results for 2005 Auction

In this section, we provide the counterfactual results for the 2005 auction. First, we find that the allocation is also highly efficient in 2005. Recall from Section 4.3 that the total annual supplying cost in the first-price CA is US$ 51.53 million. The total annual supplying cost of the minimum-cost allocation is US$ 51.49

[15]Note that a core-selecting auction may not be truthful, so in general it selects core outcomes with respect to the reported costs. In our analysis we restrict attention, however, to efficient core outcomes with respect to the truthful bids.

million over the set of relevant bid packages and US$ 50.70 million over the set of expanded package sets. This gives about 1.6% of efficiency loss in the allocation by the first-price CA.

Second, the VCG mechanism also achieves very close total procurement cost to that of the first-price CA. The total annual procurement cost under VCG is computed to be US$ 53.5 million, which is only 0.23% more expensive than the total procurement cost of US$ 53.4 million under the first-price CA. This time, the VCG payments are even closer to the core payments with respect to the truthful bids. The difference of the total procurement costs between these two points is less than 0.03% in 2005. Moreover, the individual payments are also closer; two-thirds of the nine winners receive exactly the same payments in the core point as in VCG and the rest three receive payments that are no more than 0.7% apart. Hence, in 2005, the VCG outcome is also essentially in the core.

Finally, in 2005, we have a bit larger but still small efficiency loss incurred by the allocative constraints. We consider the loss due to the market share constraints and global competition constraints. The efficiency loss in the constrained auction is 2.8% compared to the minimum-cost allocation without those constraints. In 2005 the impact of the global competition is higher; it imposes a minimum of 9 winners out of 16 bidders in 2005; in 2003 it also imposed the same minimum but out of 20 bidders.

NBER WORKING PAPER SERIES

AGE-BASED HETEROGENEITY AND PRICING REGULATION ON THE MASSACHUSETTS HEALTH INSURANCE EXCHANGE

Keith M. Marzilli Ericson
Amanda Starc

Working Paper 18089
http://www.nber.org/papers/w18089

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2012

We thank Raj Chetty, David Cutler, Mark Duggan, Jonathan Gruber, Larry Katz, Jon Kolstad, David Laibson, Ariel Pakes, Mark Pauly, Jim Rebitzer, Bob Town, and seminar participants at the University of Pennsylvania, Boston University, and the Southern Economic Association for their thoughtful comments. We acknowledge funding from the Lab for Economic Applications and Policy (LEAP) at Harvard University and the National Science Foundation. Funding was provided by the Lab for Economic Applications and Policy (LEAP) at Harvard University and the National Science Foundation. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Age-Based Heterogeneity and Pricing Regulation on the Massachusetts Health Insurance Exchange
Keith M. Marzilli Ericson and Amanda Starc
NBER Working Paper No. 18089
May 2012
JEL No. I11,I13

ABSTRACT

Little is known about consumer behavior or insurer incentives in health insurance exchanges. We analyze choice on the Massachusetts exchange, using coarse insurer pricing strategies to identify price sensitivity. We find substantial age-based heterogeneity: younger individuals are more than twice as price sensitive as older individuals. Modified community rating regulations interact with price discrimination, as our results imply higher markups on older consumers. Age-based pricing regulations would bind even conditional on perfect risk adjustment, highlighting the importance of considering insurer incentives when regulating insurance markets. Changes in age-based pricing regulation can result in transfers of 8% of the purchase price.

Keith M. Marzilli Ericson
Boston University School of Management
595 Commonwealth Avenue
Boston, MA 02115
and NBER
kericson@bu.edu

Amanda Starc
University of Pennsylvania
astarc@wharton.upenn.edu

An online appendix is available at:
http://www.nber.org/data-appendix/w18089

1 Introduction

Health insurance exchanges (HIEs) - government-run open marketplaces for private insurance - raise new questions about the effects of regulation in insurance markets. They also provide new opportunities to study consumer demand in a context with a wide range of choice. Traditionally, most individuals received either employer-based health insurance or government-provided health insurance (Medicare or Medicaid). HIEs are changing the way people purchase health insurance in that they combine retail and regulatory functions. States may set up exchanges to comply with the 2010 Patient Protection and Affordable Care Act (PPACA)[1] or as a result of their own reforms (Massachusetts and Utah). These states will have substantial latitude in designing and regulating these exchanges, and they will make choices that will shape the market for individually-purchased health insurance. However, little is known about the nature of demand for health insurance in such a setting. Understanding consumer demand and insurer incentives is important for both exchange design and for the broader regulation of insurance markets.

This paper examines consumer demand on the Massachusetts HIE, known as the Connector. The Connector provides an early look at a comprehensive HIE in action, as it was the first HIE established in the United States and has been providing coverage since 2007. Other markets also offer insight into HIEs, but have crucial differences. While Medicare Part D is like an insurance exchange, it offers a limited type of coverage (prescription drug) to a narrow age range (the elderly).[2] Employer-sponsored insurance at large employers may offer a range of choice akin to an HIE, yet it differs in regulation (e.g. plans cannot price differentially by age) and in the nature of competition (the employer negotiates directly with insurers). We add to a growing literature on choice in these contexts. Dafny, Ho, and Varela (2010) examine existing

[1] The 2010 Patient Protection and Affordable Care Act (PPACA) both mandates that all Americans carry health insurance and requires that states establish HIEs to facilitate individual purchase of health insurance. If a state does not establish an HIE, consumers will be eligible to purchase via a federal HIE.

[2] See Duggan, Healy and Scott Morton (2008) on the Medicare Part D reform, and Ericson (2012) on how consumer inertia affects the design of the Medicare Part D exchange.

employer-sponsored insurance and argue that increasing choice would lead to consumer welfare gains. In the Medicare Part D market, Abaluck and Gruber (2011) find evidence of choice inconsistencies and conclude that limiting the choice set of consumers may lead to better outcomes, while Ketcham et al. (forthcoming) argue that consumers learn over time and choose better plans.

Our analysis shows that regulation of exchanges will play a critical role in insurance markets in at least two important ways. First, regulating what counts as sufficiently generous insurance to satisfy mandates for coverage - minimum creditable coverage - will determine the plan that many people get. A majority of our sample chooses the least generous coverage sufficient to satisfy the mandate, which is less generous than coverage in typical employer plans (see also Ericson and Starc 2012). Second, regulating who pools with whom through modified community rating alters the division of surplus among consumers by limiting how insurers can vary prices based on consumer characteristics. These and other regulations are set not only by federal and state legislation, but also at the level of the HIE itself. In Massachusetts, the Connector has the power to standardize plan features and determine whether a plan provides sufficient coverage to satisfy the mandate. Exchanges also control how information about insurance plans is presented to consumers, the defaults individuals face, and the frequency of the open enrollment period.

Pure community rating, in which all consumers in a risk pool face the same price, creates a trade-off: welfare losses from adverse selection, against which are weighed welfare gains from insuring consumers against the possibility of being a bad risk by having higher expected medical spending. Modified community rating, in which insurers are allowed to vary premiums across consumers within limits, attempts to mitigate some of the welfare loss. A critical feature of modified community rating in Massachusetts has been age-based pricing: older individuals pay higher prices than younger individuals, but regulation limits the extent to which prices can vary by age. A large literature has analyzed the impact and importance of community rating laws, often considering the redistribution of risk and assuming markets are perfectly competitive.[3]

[3]For additional information, see Simon (2005), Zuckerman and Rajan (1999), Herring

Meanwhile, a recent literature (see Dafny et al. 2010, Starc 2010, and Lustig 2010 for examples) has highlighted insurer market power as a source of increasing health insurance rates. This paper knits together these literatures. Our analysis of consumer choice on the Connector provides the basis for examining how modified community rating regulation interacts with insurer incentives under imperfect competition. We emphasize the importance of accounting for both consumer behavior and strategic firm behavior when designing HIEs.

Identifying price elasticities is necessary to determine the potential for insurers to charge markups above cost (relevant for antitrust regulation), as well as the effect of subsidies for more generous insurance coverage (relevant for policy debates regarding the effect of the tax subsidy for health insurance). Existing estimates of price sensitivity for health insurance vary substantially by context, and most examine employer-sponsored health insurance.[4] The only work addressing HIEs specifically is Ericson and Starc (2012), whose measures of price sensitivity do not account for endogeneity of premiums; instead, that paper focuses on the role of heuristics in choice.

We analyze age-based pricing regulation on HIEs and show that prices by age vary not only because costs vary by age, but also because price sensitivity varies by age. We focus on the impact of this particular policy, as it is important in its own right, but also provide a general framework for modeling modified community rating in the presence of imperfect competition. Using both a reduced-form and discrete choice framework, we measure how sensitive consumers are to price. We use discontinuities in the ways firms set their prices by age to identify consumers' response to age: prices are constant within five year age blocks (e.g. 30 to 34), and then jump. In our reduced-form specifications, we find that total spending rises nearly one-for-one with price increases. However, this misleadingly suggests little consumer response to price; in fact, because consumers are already clustered at the cheapest plans (25% choose the cheapest plan available to them in our sample period), they have little ability

and Pauly (2006), and Finkelstein, Poterba, and Rothschild (2009).

[4]See Carlin and Town (2007), Gruber and Washington (2006), Bundorf, Levin and Mahoney (2008) and Einav, Finkelstein and Cullen (2010).

to reduce insurance spending in response to price increases. A discrete choice model is therefore more appropriate in measuring the value that consumers place on plans, as derived from their underlying demand for medical services.[5] Plans with the same level of coverage are not typically perfect substitutes, as insurers have different networks of providers; consumers may also attach value to firm brand and reputation. Moreover, individuals are constrained by their choice set.

We identify substantial variation in demand elasticities: younger consumers are significantly more price sensitive than their older counterparts. Furthermore, there is heterogeneity within age groups. The estimates show that consumers in the 75th percentile of price sensitivity have elasticities four times larger than their counterparts in the 25th percentile. Heterogeneity between and within observable groups is important for evaluating potential policies, such as the minimum creditable coverage regulation and age-based pricing regulation.

Insurers may price discriminate by responding to heterogeneity in consumer demand that is correlated with observable tags. (Here we study age, but other contexts could include gender, geographic location, and race.) Using our estimates, we simulate the ability of insurers to price discriminate under various models of market structure and highlight the potential welfare effects of alternative pricing regulations. In order for insurers to price discriminate, there must be consumer heterogeneity in preferences that can be both observed and priced. We conduct a counterfactual exercise in which we simulate premiums and estimate the distributional consequences of eliminating or tightening age rating rules. Ultimately, the disparities in price sensitivity by age imply that age rating rules are one of the most important regulatory features of the exchange.[6] We estimate the effect of moving from no restrictions on age-based

[5] Moreover, consumers may be daunted by the complexity of the insurance product and difficulty forcasting their future medical expenses. In such cases, consumers may rely on heuristics. We explore this idea in further detail in Ericson and Starc (2012).

[6] Geruso (2011) considers the impact that preference heterogeneity has on welfare in a model in which insurers are perfectly competitive. Our model, by contrast, shows how imperfectly competitive insurers can amplify this effect and how differences in preferences can lead to transfers in the absence of cost differences via price discrimination.

pricing to regulations that prohibit age-based pricing. Based on insurer price discrimination alone, this change leads to transfers of 8% of the purchase price of insurance; accounting for cost differences between ages leads to even larger price increases.[7]

Finally, we allow consumers to opt out of the market to capture the importance of the individual mandate. Even in the case of full risk-adjustment, so that costs do not differ by age, differences in preferences alone can lead to this market unraveling. We allow consumers to opt out of the market completely and the resulting change in the composition of consumers in the market to affect markups. If younger consumers are allowed (in the model or by law) to opt out of coverage, modified community rating can lead them to opt out of coverage even absent cost differences. As the most price-sensitive consumers opt out, less price-sensitive consumers are left in the market, leading to higher markups. This, in turn, leads more price-sensitive consumers to opt out of the market until there are only price insensitive consumers left in the market.

This paper shows how the heterogeneity in consumer preferences for insurance products noted in the literature (Cohen and Einav 2003, Cutler, Finkelstein, and McGarry 2006) interacts with public policy. We identify the regulations that are of key importance in this market and explain why these regulations are critical given consumer demand and the strategic reaction of insurers to both demand and regulation. Ultimately, choosing the set of consumers who form a risk pool is critical for determining the allocation of surplus in insurance markets. We highlight the importance of considering differences in preferences as well as differences in costs. As states decide how to define their exchanges and which segments of consumers to include - such subsidized enrollees, younger consumers eligible for catastrophic plans, or employees in small groups - understanding differences in preferences is just as critical as understanding differences in risk.

The paper proceeds as follows. The second section describes the Massa-

[7] These implications - though a transfer, rather than a welfare loss - are larger than the welfare loss from selection (see Bundorf, Levin, and Mahoney 2008 and Einav, Finkelstein, and Cullen 2010) and highlight the importance of considering the incentives of imperfectly competitive insurers when designing and regulating insurance markets.

chusetts Connector, rating regulation, and some reduced-form results. Section 3 describes reduced-form evidence of consumer spending elasticity, while Section 4 expands on this analysis to incorporate a discrete choice approach. Section 5 discusses incentives for non-uniform pricing and describes the related counterfactual exercise. Section 6 concludes.

2 The Massachusetts Connector: Context and Data

2.1 Massachusetts' Health Reform

The state of Massachusetts signed its health care bill into law in April 2006, with the goal of providing universal coverage for its residents; the reform, in turn, served as a national model for health reform. This reform had many features, including expansions in public coverage, and individual and employer mandates. A key feature of this reform was the individual mandate, which required all Massachusetts residents to purchase a minimal level of health insurance coverage (minimum creditable coverage) or face a penalty equal to half of the premium of the lowest cost health insurance plan (for their age) offered through the exchange. To facilitate consumers purchasing insurance, the state required employers with 11 or more employees to make a fair and reasonable contribution to employees' health insurance costs. It also established the Commonwealth Care program, which provided free or subsidized coverage to lower income residents, who earned up to 300% of the federal poverty level.

Finally, the reform established an unsubsidized health insurance exchange (the Commonwealth Choice program, run by the Connector) to facilitate non-group coverage purchased directly by households and small group purchase of insurance. The Commonwealth Connector Authority operates as a quasi-public agency and has offered health insurance through the Connector since May 1, 2007 (with the mandate taking effect July 1, 2007). The Connector shapes the market for individual coverage in Massachusetts in a number

of ways. It operates the exchange's website[8] and chooses which features of insurance plans are highlighted.

The Massachusetts reform has been effective at reducing the number of uninsured individuals. In 2009, 97.3% of the population was insured (Long and Phadera 2009), with increases in the insured coming from individuals purchasing insurance through the Connector, through increased offering of employer-provided health insurance, and through expansions in subsidized coverage (Gruber 2011). Kolstad and Kowalski (2010) show that the Massachusetts reform not only increased coverage, but also decreased hospitalization for preventable conditions. However, the effect of the health reform and the Connector on the level and growth rate of premiums is a point of contention. How the Connector affects insurance prices depends on both consumer demand and market structure. By characterizing both in this paper, we provide a foundation for future analysis of the impact of the Connector.

2.2 Regulation of the Health Insurance Exchange

There are two important regulations in the market:

Minimum creditable coverage (MCC): MCC is the least generous plan that is sufficient to comply with the mandate. The Connector is responsible for determining MCC for the state based on a combination of actuarial value, out-of-pocket maximum, deductibles, covered physician visits, and prescription coverage. In Massachusetts, MCC includes prescription drug coverage and three check-ups, caps deductibles at $2000 for an individual and $4000 for a family, and caps out-of-pocket expenditures at $5000 for an individual and $10,000 for a family.

A large number of policies just satisfying MCC are available, and they are quite popular. Therefore, regulation regarding the definition of MCC is likely to be critically important in a market with a mandate.[9] While MCC may

[8] The website is http://www.mahealthconnector.org.

[9] Finkelstein (2004) finds that minimum standards can reduce enrollment by potentially exacerbating adverse selection. However, in the presence of a mandate, such concerns are much less pressing.

not be directly under the control of an exchange regulator or designer, it will dramatically shape the market within the exchange.

Modified Community Rating: Modified community rating rules apply to pricing on the exchange. Specifically, rates for the same product have to fall within a 2:1 band across ages and geography. For a given plan, the highest quoted premium can be at most twice the lowest quoted premium. In addition, no medical underwriting is allowed, and plans are guaranteed issue (no one can be denied coverage). These rating rules are critical in shaping premiums in the market. Age, in particular is a critical feature of rating.

Figure 2 shows that insurers are clearly constrained by regulations for age-based pricing: the average monthly premium for a 27-year-olds is just over $300, which the premiums for older consumers are just over $600. The choice of rating-bands will alter the division of surplus among young consumers, older consumers, and firms. While PPACA specifies a 3:1 maximum allowable age rating band in the individual health insurance exchanges, states can impose more strict regulation. For example, Maryland has chosen a ratio of 2.8 to 1 as a rating band (Carey and Gruber 2010).

2.3 Making Choices on the Exchange

The exchange offers a variety of health plans administered by the major private insurers in the state.[10] Insurers initially had relatively wide latitude in designing these plans, which were grouped into tiers based on actuarial value: bronze, silver, and gold. Bronze plans are generally less generous (higher cost-sharing) and therefore tend to be cheaper. Gold plans are the most generous and hence most expensive, while silver plans forge a middle ground. Beginning in 2010, the Connector required plans to take one of six standardized forms (bronze low, bronze high, etc.), though plans may still differentiate themselves based on their provider networks. In addition to this main market, there is a separate market for young adult consumers aged 18 to 26, in which plans tend

[10] In our sample, the following firms sold insurance via the connector: Blue Cross Blue Shield of Massachusetts, Fallon Community Health Plan, Harvard Pilgrim Health Care, Health New England, Neighborhood Health Plan, and Tufts Health Plan.

to have more limited coverage, such as optional prescription drug coverage.

Consumers face a number of steps when purchasing insurance from the Connector. (Screenshots from the purchasing process are included in the Appendix.) After entering demographic information, consumers are offered a choice of plans that vary along a number of dimensions, including copayments, deductibles, and premiums. Importantly, the plans are placed into tiers; this grouping might affect consumer choice. Prior to the 2010 standardization of plan types, consumers needed to weigh multiple dimensions of plans (copayments, coinsurance, dental coverage), even within each tier. Finally, the consumers enroll.

The website itself, in addition to regulation, has the potential to shape consumer choices. Tiering can also affect how insurers design plans; for example, they may design plans to meet the minimum level of generosity in a tier. The way information is presented, plan features highlighted and the order in which plans are sorted may also affect consumer behavior. For instance, Ericson and Starc (2012) finds a discontinuity in preference for the minimum choice plan. During the initial period, the website sorted plans according to price (as opposed to, for example, consumer satisfaction) so consumers may have inferred that price was the most important variable differentiating these plans.

2.4 Data and Descriptive Statistics

We use transaction-level data (purchase, cancellation, and payments) from the unsubsidized market (Commonwealth Choice) from the beginning of the Connector's existence in July 2007 until December 2009. We observe approximately 50,000 transactions. There are large spikes in initial enrollment during the first month of the Connector's existence as well as just before the individual mandate's financial penalties took effect in December 2007, with a steady-state enrollment of approximately 1,000 households per month. Appendix Figure A.1 plots a histogram of the number of individuals choosing single coverage joining the Connector for the first time, by month (the majority of purchases are for single coverage).

Table 1 describes the demographics of these consumers: most are young, with an average age of approximately 35. Most purchase individual, rather than family plans, and a sizable percentage lives in Middlesex County (which includes Boston suburbs like Cambridge and Somerville). The average premium paid is about $420.30 per month, but varies substantially by age. We also examine how long consumers stay in the Connector since their initial enrollment. Figure 1 gives the Kaplan-Meier survival curve for time spent in the connector, split by tier of plan chosen. Approximately half of our observations are censored because these individuals are still enrolled in insurance when our data sample ends. The median consumer is enrolled in a Connector plan for about 13 months, and there is no spike in individuals exiting the Connector after their one-year contract is complete. It shows only small differences in enrollment duration between tiers.

This paper focuses on consumers purchasing individual coverage (as opposed to household coverage), since the majority of plans sold are of this type. We exclude consumers eligible for young adult insurance aged 26 and under from this sample because they have a different choice set and rarely purchase plans other than young adult insurance plans. Our choice analyses focus on two subsets of the data: November-December 2009, and July-December 2009. Because we observe transaction-level data, we do not observe all the plan prices that individuals face. However, for November and December 2009, we collected an extensive set of price quotes from the Connector website using a Perl script. Because the plan menu is relatively constant from July through December, we are able to infer the prices individuals faced from July to October with a high degree of accuracy. The Data Appendix gives more details. The choice of sample period does not have a strong effect on the results, and we show the robustness of our results to various sample selections.

3 Consumer Response to Price: Identification Strategy and Reduced-Form Evidence

3.1 Coarse Firm Pricing

The level of consumer price sensitivity is an important feature of demand that affects both policy design and insurer price-setting behavior. Estimates of price sensitivity are difficult to identify because unobserved plan characteristics may be correlated with price. The results from the literature are mixed: some studies, such as Cutler and Reber (1998) find relatively high elasticities among young, healthy consumers, but other studies find that demand for health insurance is typically inelastic.[11] We use a regression discontinuity identification strategy based on coarse pricing rules used by firms to identify the effect of price on choice. By law, firms may vary prices (within broad limits) by both zipcode and age. However, firms price more coarsely than the zipcode and age level. We first use coarse geographical pricing to define markets. Instead of varying prices for each zipcode, firms set prices for larger geographic regions that roughly correspond to hospital referral networks that may be a good proxy for underlying costs.[12] For example, Blue Cross Blue Shield charges three sets of premiums: one set for western Massachusetts, one set for the greater Boston area, and one set for Cape Cod. We use this variation to define a geographical region that is a set of zipcodes in which prices do not vary within a plan-age cell. (See Data Appendix for details.)

We use coarse age-based pricing rules to identify price sensitivity. Firms do not vary prices continuously as individuals age. While prices have discrete jumps at various ages, preferences are likely to evolve continuously across ages. The listed premium given plan has discrete changes at ages ending in 0 or 5 (30, 35, 40, etc.). Figure 2 shows jumps at each age in the average premium for a constant set of plans.[13] These jumps translate into very similar consumers

[11] See the survey evidence of Kreuger and Kuziemko (2011).

[12] These differences could also be driven by differences in medical utilization.

[13] The marginal cost of choosing a more generous plan jumps correspondingly, as shown by Appendix Figure A.2. The ratio of the cost of the average gold plan to that of the average

facing very different vectors of premiums: the underlying preferences of a 39-year-old and 40-year-old will likely be very similar; however, they will face different premiums for the same plan.

Our identification strategy relies on the discontinuity in price being unrelated to demand; i.e. preferences evolve continuously as an individual ages, so turning 40 is like turning 39 or 38. Based on our conversations with insurance firms, these discrete jumps in price result from firms' menu costs when setting premiums. The combination of age bins and zipcodes alone gives rise to over 40,000 potential prices, all of which must be submitted to a regulator for approval. Optimally pricing each combination and submitting it for approval would be a costly exercise for firms and could trigger adverse regulatory action. Moreover, Chu, Leslie, and Sorenson (2011) show that firms can obtain profits close to a perfectly price discriminating firm using coarse pricing rules. Finally, regulators have identified these discontinuities as a potential problem and have introduced legislation to "smooth" the relationship between premiums and age, suggesting they are not a result of shifts in utilization or preference.

While firms could price in discrete age blocks if the cost of an insured individual changed dramatically at each age cutoff, this alternative explanation for the jump in prices is not supported by the data. While diagnostic tests (such as mammograms) are recommended for patients beginning at the age cutoffs, observed medical spending in the Medical Expenditure Panel Survey (MEPS) rises smoothly and shows no systematic discontinuities in health expenditures at round numbered ages. Thus, differences in spending are unlikely to account for such large price jumps.

Appendix Table A.2 supports our identification strategy by showing that characteristics of enrollees' zipcodes do not change discontinuously between age categories, with the exception that enrollees over age 55 seem to be slightly more wealthy, employed, and white. This may lead us to slightly underestimate the price sensitivity of this age category. Similarly, the density of individuals enrolling in the Connector does not change at the various age cutoffs. Fig-

bronze plan varies slightly within each age category but stays between 1.8 and 2.

ure A.3 shows the number of enrollees in each one-year age bin (we do not have exact birthdate, only age in years). Visual inspection indicates there is no general pattern of densities dropping at round numbered ages, with perhaps an anomalous low enrollment for individuals aged exactly 50 years. The final column of Appendix Table A.2 shows that the density doesn't change discontinuously at any breakpoints, with the potential exception of the age 50 breakpoint.[14]

3.2 Reduced-Form Evidence on Response to Price

This section examines how choice of health insurance responds to price, using reduced form evidence on total spending. When the marginal price of more generous insurance plans decreases, consumers may chose to increase the generosity their health plan. One way to summarize the response to price is the insurance spending elasticity, which relates relate total premiums paid to the list prices individuals face. We summarize the effect of a price change on insurance spending using the following model:

$$\ln y_i = \eta \ln (p_i) + \gamma \omega_i,$$

where y_i is the total insurance premiums paid by individual i (given the actual prices), p_i is a price index for a representative bundle of plans, and ω_i is a vector of individual characteristics. The insurance spending elasticity is given by η and says that if the price index rises by 1%, the total spending rises by η%.[15] If $\eta < 1$, individuals respond to higher prices by reducing their spending on insurance, while if choice of insurance plan stayed the same, then $\eta = 1$.

In this context, the percentage price increase at each threshold varies

[14] The table presents results using the November-December Analysis Sample; the results using the July-December sample are similar and, in fact, do not contain any significant differences in zipcode characteristics at age 55.

[15] Of course, while identifying η is a valuable way of summarizing the data that can facilitate out-of-context prediction, individuals do not in fact face a continuous choice of dollars spent on health insurance; the discrete choice individuals actually face is modeled in Section 4.

among plans and insurers.[16] We therefore create a price index, in which each plan is assigned a weight. Because the plan menu varies by geographic region, we create geographic-specific weights: a plan's weight is the fraction of people in a geographic region who chose that plan, averaged over July to December 2009.[17] Column 1 of Table 2 shows how the price index jumps at each age threshold. It presents the results of the following regression:

$$\ln p_i = G(a) + \sum_{s \in 1,...7} 1_{a \geq a_s^*} \pi_s + \gamma \omega_i, \tag{1}$$

where $G(a)$ is a linear spline in age and ω_i includes gender, month of enrollment, and indicators for geographic region.[18] The coefficients π_s multiply indicator variables for whether age is greater than or equal to each of the age thresholds (each value of a_s^*) used for pricing. Each value of π_s shows how the price index jumps at the threshold a_s^*: for instance, we see that the price index increases by 20.4 log points when an individual turns 50. The jumps in prices are relatively small at age 30 and 35 but are more substantial at older ages.[19]

Next, we examine how total spending on premiums changes at each age threshold shown in Column 2 of Table 2. It presents the results of the regression

$$\ln y_i = G(a) + \sum_{s \in 1,...7} 1_{a \geq a_s^*} \kappa_s + \gamma \omega_i \tag{2}$$

where $G(a)$, ω_i, and the age category indicators are the same as in Equation 1. The values of κ_s from this regression show how spending on premiums jumps

[16] By contrast, a change in tax deduction for employer-sponsored health insurance (as in Gruber and Washington 2005) would lead to the same percentage change in price for all the plans, eliminating the need to construct a price index.

[17] To construct a reasonable price index, we exclude geographic regions that had fewer than 10 zipcodes, as well as geographic regions that had fewer than four insurers.

[18] Because we know the pricing model, gender and the linear age spline do not predict prices; they are included for comparability with later regressions.

[19] The measured increases in the price index at the age thresholds (and their relative magnitudes) vary depending on how the price index is constructed. We gave plans weights based on popularity in a geographical region. Ideally, we would assign age-group specific weights (to create a Laspeyres or Paasche index). However, the sparsity of the data makes this an unappealing route. The construction of the price index has no bearing on how we measure the change in spending at each threshold (but is relevant for estimating η).

at each age discontinuity. Thus, we see spending on premiums jump 19.2 log points at age 55, controlling for linear age trends above and below 55, along with other variables.

Comparing the percentage increase in spending (κ_s in Column 2) to the percentage increase in the price index (π_s from Column 1), we see that the increase in spending is slightly less than the increase in prices for all age thresholds except age 55. The bottom panel of Table 2 shows the results of an instrumental variable regression that instruments for price index by age-category, controlling for an age spline.[20] Column 1 is thus the first-stage of this IV regression, and the F-statistic for excluded instruments (age-discontinuities) is a substantial 2823. The resulting estimate of η is 0.962 (s.e. 0.176), indicating that a 10% increase in the price index leads to a 9.6% increase in total spending in this population - a relatively limited response by individuals.[21]

The pattern of results above is robust to larger samples. We limited our sample to the July through December 2009 sample, because outside that time period we do not have the full menu of prices an individual faced. Nonetheless, we can run an (imperfect) analog of equation 2, where the unit of observation is a plan chosen by an individual and the dependent variable is the price paid. Controlling for plan fixed effects, we can measure how the price paid for a fixed bundle of plans changes across age thresholds. Columns 1 and 2 of Appendix Table A.1 runs these regressions for the July-Dec. 2009 sample and the full sample. Results are similar; the sole exception is the price increase at age 40, which is estimated to be 14-15% in the July to December sample, but is only 9% in the full sample. Columns 3 and 4 estimate equation 2 on both samples and show similar results.

This nonstructural approach shows that there is little response in individual choice to price increases: the increase in spending is approximately equal to the increase in prices. Yet Section 4 will show that these results do *not* imply that individuals are not sensitive to price. Rather, individuals are al-

[20] Linear age splines have knots at each age threshold. Additional controls include indicators for month, geographic region, and gender.

[21] An age group-specific insurance spending elasticity could be estimated by as $\eta_s = \frac{\kappa_s}{\pi_s}$, but the standard errors on each η_s would be extremely large.

ready gravitating to the least generous tier and cheapest plans available. Thus, despite the wide range of plan generosities available in the Connector, individuals do not have much latitude to respond to a price increase. They are simply unable to substitute to cheaper plans. Thus, these results highlight the importance of context in determining the effect of policy changes (e.g., altering the tax exclusion for employer-provided health insurance) and motivate our structural model of consumer preferences in Section 4.

4 Discrete-Choice Model

4.1 Theoretical Model

We now explicitly model consumers' discrete choice of insurance plan using a standard logit model. We assume that consumer i's utility of plan j is given by:

$$u_{ij} = \delta_j + \mu_{ij} + \xi_{ij} + \varepsilon_{ij},$$

where δ_j is the mean utility of a plan, μ_{ij} represents the (mean-zero) component of a plan's utility that varies based on observed individual characteristics (e.g., age or location), and ε_{ij} is an error term that is independently and identically distribution (i.i.d.) extreme value. This implies shares can be written as:

$$s_{ij} = \frac{\exp\left(\delta_j + \mu_{ij} + \xi_{ij}\right)}{1 + \Sigma_j \exp\left(\delta_j + \mu_{ij} + \xi_{ij}\right)},$$

where s_{ij} represents the probability that consumer i purchases product j. In the absence of individual heterogeneity μ_{ij}, the δ_j parameters simply represent an inversion of the observed market shares for each plan. The mean utility δ_j can be a plan fixed effect, or can be a function of plan characteristics X_j, such as insurer (brand), price, deductibles, and copayments. A given insurance plan (e.g., HMO Blue Basic Value) is offered in multiple markets.

As discussed in the previous section, we identify the premium coefficient under the assumption that preferences evolve continuously with age, so that discontinuities in mean utilities at round-numbered ages are solely attributable

to discontinuous changes in premiums.[22] More formally, let δ_{j30} be the mean utility of product j offered to consumers who are age 30, where $\delta_{j30} = X_j'\beta + \alpha p_{j30}$, and X is a vector of plan fixed effects. Similarly, let δ_{j29} be the mean utility of product j offered to a consumer who is age 29. Our choice model gives us consistent estimates of both utilities. Then in the absence of age trends, the price coefficient can be simply written as:

$$\alpha = E_j \left[\frac{\delta_{j30} - \delta_{j29}}{p_{j30} - p_{j29}} \right].$$

Of course, there are age trends in preferences. We allow for δ_j to evolve continuously over ages, but limited data require that we place some structure on how it does so. Estimating a separate linear spline in age (with six knots) for each of the 21 different plans would allow maximal flexibility but is infeasible in our data. We allow for preferences for plan tier (bronze, silver, gold) to evolve flexibly with age. Further, we allow different plans within a tier to have different qualities. Ultimately, the assumption that identifies the price coefficient in these specifications is that age-specific deviations in preference for plans within a given tier are not correlated with prices. This seems a reasonable assumption, and our results do not change substantially when we allow for more variability in age-specific preference for plans.

4.2 Estimation

The model is estimated using a conditional logit approach in Table 3. This allows us to interact consumer characteristics, such as age, with plan characteristics, such as price. A critical feature of demand in this market is consumer heterogeneity in preferences. Panel A, Column 1 estimates the model without allowing for consumer heterogeneity, while column 2 allows price sensitivity to vary linearly by age. The data strongly reject constant price sensitivity by

[22] Without assuming that preferences evolve smoothly, the premium coefficient is not separately, nonparametrically identified from variation in preferences over plans. This is because the premium is itself a (highly nonlinear) function of demographic characteristics, such as age, that may also impact preference for plans or benefit designs. This is why allowing flexible preferences over plans is critical in Table 2.

age. Furthermore, by comparing columns 2 and 3 we see that accounting for variation in preferences is important to estimating the level of price sensitivity as well. Specifically, the unobserved product characteristic, denoted above by ξ_{ij}, consists of age and geographic specific deviations from mean plan quality - the added utility that a specific plan brings to an older consumer, for example. These are likely to be positively correlated with price, and failing to account for such effects will bias the price sensitivity toward zero.[23] The estimates in column 3 indicate that the oldest consumer in our sample (64) is roughly half as price sensitive as the youngest consumer in our sample.

Alternative specifications confirm the pattern of lower price sensitivity by age. Panel B of Table 3 separately estimates the model by five-year age bands. These results for each 10-year age bin still show the variation in price sensitivity by age, indicating our structural assumptions are not too restrictive. Finally, for use in some counterfactual exercises, Panel C of Table 3 cuts the age span in the Connector in half and runs the model separately for those under age 45 and age 45 and older. The younger consumers are substantially more price sensitive, and the difference is statistically significant. Furthermore, the results indicate semi-elasticities (which describe the percent change in enrollment given a $100 increase in premiums) for younger consumers of around -3 and for older consumers of just above -1. Figure 3 maps out these price elasticities under various specifications, highlighting the pattern in semi-elasticities over the life cycle.

Table 4 shows additional specifications. The first two specifications break out price sensitivity into five-year age bands. Column 2, in particular, shows that the trend may not be linear, but that the oldest consumers have substantially different preferences than their younger counterparts. Column 3 estimates a nested logit, in which consumers first choose a plan tier (gold, silver, or bronze) and then choose a policy from within that tier. The dissimilarity parameter is an inverse measure of the correlation between error terms in each nest, and a dissimilarity parameter of one would indicate the

[23] Furthermore, later in this section, we will discuss how to use the identifying variation from coarse insurer pricing.

model collapses to the conditional logit. The low dissimilarity parameter for bronze and silver plans indicates that consumers see these plans as fairly close substitutes. However, a dissimilarity parameter near 1, gold plans are not close substitutes, indicating that networks and brand name factor highly in the decisions of consumers who are likely to purchase gold plans. Column 4 includes mixed logit results, in which the price sensitivity is allowed to take on a log-normal distribution, shifted by age category. The results show that distribution is shifted toward zero (less price sensitive) for older consumers. Nonetheless, because of data limitations and the flexibility the specifications provide, these results are somewhat noisy.

Taken together, the results from Tables 3 and 4 are striking. The elasticities of the youngest group, those 27 to 35, are nearly twice as large in magnitude as those of the oldest group. The raw data driving these results can be summarized as follows: For older consumers, the marginal cost of gold plans relative to bronze plans is much higher than for younger consumers (Appendix Figure A.2). Despite these differences, the fraction of consumers purchasing bronze plans stays relatively flat with age. This indicates that older consumers have a lower distaste for price, a higher preference for more generous coverage, or both.

Various demographic factors could be driving the preference heterogeneity we see in the data. For the pricing exercise in the next section, it does not necessarily matter whether age is simply a signal for another demographic factor correlated with preferences or not. We note that younger consumers are not from lower income zipcodes in our data. However, because older individuals are more likely to be married, the selection of older consumers into the exchange may differ, as some married consumers have access to insurance through a spouse. In addition, older consumers are less likely to report that they are in excellent health; Strombom et al. (2002) report that older and sicker consumers tend to be less price sensitive. Finally, the relatively older consumers in our sample might be more financially sophisticated, leading them to more heavily weigh characteristics other than price when making decisions.

5 Application: Age-Based Pricing Regulation

5.1 Motivation

In this section, we model how age-based pricing regulation affects markets in the presence of imperfect competition. Existing work (e.g., Blumberg, Buettgens, and Garrett 2009) has assumed that insurers price differentially by age solely due to cost differentials, so age-based pricing regulations only bind to the extent the ratio of costs exceeds the maximum allowable ratio of prices by age. However, Section 4 shows that price-sensitivity varies by age, and so insurers would want to price discriminate and charge higher prices to older enrollees, even if health costs did not vary by age. This section analyzes the impact of age-based pricing regulations in the presence of age-based heterogeneity in price sensitivity. We develop the model in the context of age, but the same logic would apply to any observable tag by which costs and preferences varied.

We analyze the effect of three types of age-based pricing regulations:

- Age-Pooling: firms cannot vary prices by age.
- Age-Bands: firms can vary prices by age, but the ratio of the highest price to lowest price cannot exceed θ.
- Age-Unconstrained: firms can vary prices by age.

In all cases, we assume that if a plan is offered, it must be offered to all ages. Note further that pooling and unconstrained prices are simply special cases of age bands (where θ =1 and ∞, respectively). On the Massachusetts Connector, $\theta = 2$, while PPACA requires states to set $\theta \leq 3$.

5.2 Model and Theoretical Predictions

Consider two types of consumers, old and young, who are purchasing an insurance plan.[24] Costs rise with age, so the old have an average cost (to the

[24]Here, we assume both groups have a high enough willingness to pay to purchase insurance, so that selection out of the market is not an issue. That is, we assume the mandate

insurer) of c_H, greater than the cost of the young c_L. Let fraction σ of the population be old, and fraction $1-\sigma$ be young. There are $N \geq 2$ profit-maximizing insurers, each offering a single plan[25] that is available to the young and the old. Insurers can determine whether an individual is old or young, but cannot further determine the expected cost of the individual. Hence, each insurer can set two prices, one for old and one for young individuals: p_H and p_L.

We first examine how regulation affects pricing in perfectly competitive markets, in which products are identical and firms make zero profits.[26] Stricter limitations on age-based pricing (reducing θ) transfers resources from old individuals to young individuals. Age-bands are only binding up to the ratio of costs between the two groups. Prices are summarized below:

- Under the Age-Pooling regulation, prices are equal to population average cost: $\bar{p} = \sigma c_H + (1-\sigma)\, c_L$.
- Under the Age-Unconstrained regulation, prices are equal to each type's average cost: $p_H = c_H$ and $p_L = c_L$.
- Under the Age-Bands regulation with $\theta \leq \frac{c_H}{c_L}$, prices for the young are above their cost, and for the old are below their cost: $p_H^* = \theta p_L^*$ and $p_L^* = \frac{1}{(1-(1-\theta)\sigma)} \left[\sigma c_H + (1-\sigma)\, c_L\right]$. When $\theta > \frac{c_H}{c_L}$, the regulation does not bind, and so $p_H = c_H$ and $p_L = c_L$.

However, in an imperfectly competitive market, the prices set by insurers for each group are determined not only by costs, but also by that group's elasticity of demand. Hence, prices for old and young consumers may differ due to *price discrimination motive* as well as a *cost differential motive.* Thus, we must consider how characteristics other than cost affect prices when modeling age-based pricing regulations. The existence of the price discrimination motive means that low-cost consumers could face higher prices than high-risk consumers if they are not sensitive to the price of insurance policies

is effective. Section 6 examines noncompliance with the mandate.

[25] We abstract away from adverse selection between policies of different quality.

[26] Formally, let there be a continuum of consumers normalized to measure 1. When multiple firms offer a plan at the same price, consumers are evenly distributed across the firms.

(the "worried-well"; see Starc 2012). However, it may also be the case that the higher prices high-risk consumers face are amplified by low price sensitivities.

Now let the market be imperfectly competitive, with all plans being identical in average quality. Let $\tilde{s}_{ja}$ reflect the share of age group a that purchases insurance at firm j, and let $s_{jH} = \sigma \tilde{s}_{jH}$ and $s_{jL} = (1-\sigma)\,\tilde{s}_{jL}$ be the number of each group purchasing insurance at firm j. Then, we can write the profits of firm j as:

$$\Pi_j = s_{jH}\left(p_{jH} - c_H\right) + s_{jL}\left(p_{jL} - c_L\right).$$

Firms set prices based on their first-order conditions (which we assume are unique), subject to the age-based pricing regulations they face. We drop the j subscripts below. We define a few terms: let s_H and s_L be functions of p_H and p_L, respectively, so that s_i' gives the change in type i's enrollment as p_i changes. Let total enrollment be $S = s_H + s_L$. For use in the Age-Bands pricing, define weighted demand $\bar{S} = \theta s_H + s_L$. When the bands are binding, write p_H as an implicit function of p_L, and $\bar{S}$ as a function of p_L, so that $\frac{d\bar{S}}{dp_L} = \theta^2 s_H' + s_L'$.

Proposition 1 *Assume markets are imperfectly competitive. Then, under the Age-Pooling regime,* $p^{Pool} = \frac{1}{\frac{dS}{dp}}\left(s_H' c_H + s_L' c_L\right) - \frac{S}{\frac{dS}{dp}}$. *Under Age-Unconstrained,* $p_H^{Un} = c_H - \frac{s_H}{s_H'}$ *and* $p_L^{Un} = c_L - \frac{s_L}{s_L'}$ *. If Age-Bands are binding,* $p_L^{Band} = \left(\frac{\theta s_H'}{\frac{dS}{dp_L}} c_H + \frac{s_L'}{\frac{dS}{dp_L}} c_L\right) - \frac{\bar{S}}{\frac{dS}{dp_L}}$ *and* $p_H^{Band} = \theta p_L^{Band}$.

Proof. Immediate from first-order condition. ■

Proposition 1 shows that under the Age-Unconstrained policy, firms simply set prices for each group equal to cost plus a markup inversely proportional to the elasticity of that group's demand. An insurer can only set one price under Age-Pooling, which is equal to a markup term inversely related to the elasticity of population demand, plus a cost term, where the relative weight on each cost term is that groups' share of the marginal change in demand. The optimal price under binding Age-Bands is similar to that under Age-Pooling, except the markup term is now inversely related to weighted demand $\bar{S}$, and the weight on each cost term is given by θ. The first-order condition thus

takes into account that the price for the high-cost group is θ times that for the low-cost group. If the low-cost group (young) is more price sensitive than the high-cost group (old), there are two reasons for the high-cost group to prefer a pooling or pseudo-pooling arrangement. First, as always, more low-risk types lower the average cost. However, more price-sensitive individuals also lower the optimal markup of the insurer. We use these first-order conditions for price setting in the counterfactual exercise that follows.

5.3 Counterfactual Exercise: Changing Age-Based Pricing Regulation

In this section, we examine how alternative age-based pricing regulations would affect prices and welfare on the Massachusetts Connector. We use the stylized pricing rule developed in Proposition 1, but we do not deny that firm strategies may in fact be much more complicated.[27] However, our framework matches how firms set prices on the Connector in practice. Insurers submit quotes to the Connector, which consist of a base rate and an adjustment factor. This adjustment factor takes into account any differences in costs and may account for differences in consumer preferences.

We consider two types of counterfactual scenarios. In the first one, firms price to two age groups (over/under 45), and in the second, firms price to three age groups (27-35, 35-45, 45+). Before examining age-based differences in costs, we simulate how insurers would price in a stylized environment where a risk-adjustment scheme perfectly compensated insurers for age-based differences in costs. (There is no risk adjustment in the Connector.) When firms are unconstrained by age-based pricing regulation (Age-Unconstrained), young consumers are sufficiently price sensitive to ensure price competition on their policies. Our estimates indicate that the optimal markup for consumers under 45 is 10%. However, because there is a sharp drop in price sensitivity at age 45, the older group of consumers faces high markups (roughly a third of

[27] For instance, we do not capture competition over multiple, related products. For example, a firm may target a specific group of consumers with one plan (i.e., young individuals or families with a bronze plan), and another group with a different plan.

price) in the absence of regulation. Now consider the other extreme: Age-Pooling, so firms can set only one price for all ages. The presence of younger consumers in a pool with older consumers can partially hold down markups, and because they are relatively more prevalent in the population, their higher elasticity has a large weight on the average markup, which is approximately 20%. Thus, moving from Age-Unconstrained pricing to Age-Pooling would entail substantial redistribution away from younger individuals.

These simulations show that older consumers would be willing to pay younger consumers to face premiums set for the entire population, even if risk adjustment perfectly compensated insurers for age-related differences in costs.[28] Table 5 shows the transfers from younger individuals that would result from moving from Age-Unconstrained to Age-Pooling or Age-Bands. When firms price to two age groups, pooling leads to a 6% average increase in premium for those under 45, while if they price to three groups the transfer increases to 7.6% (column 2).[29] As the age-bands are relaxed, the transfers away from the under-45 year olds is correspondingly reduce. Note that these transfer estimates *assume no cost differences* between the two groups; any cost difference would exacerbate this transfer from younger consumers to older consumers.

We next examine differences in costs by age-group. Table 6 uses data from

[28] The desire to face younger consumers' prices arises because they are more price-sensitive and does not rely on younger consumers having lower costs. Note that there could also be different optimal markups over classes of products as well. For example, if consumers purchasing bronze plans are more price-sensitive than average, the average markup on these plans will be lower. In addition, the minimum price effect (if it is truly a heuristic) will induce plans to compete vigorously to be the cheapest. None of these effects are explicitly modeled here, however additional specifications (not presented) indicate that the price sensitivity for bronze plans is significantly higher than their silver or gold counterparts.

Note that a dollar reduction in premium to a younger consumer is more valuable in utility terms than the corresponding dollar increase in premium to an older consumer. The only potential welfare gains come from more efficient sorting of consumers. We abstract from consumer reoptimization.

[29] The simulation results are somewhat sensitive to a number of choices, including which variation to use (for example, identifying the price sensitivity of a 31-year-old using the price jump at 30 or the price jump at 35), the flexibility of the demand specification, and the weighting placed on different age groups. Therefore, the results in each simulation are always similar in direction and order of magnitude, though not identical.

the 2008 Medical Expenditure Panel Survey (MEPS) on the health costs of difference groups. To construct the table, we restrict the sample to individuals 27-64 and with moderate to high incomes and private insurance, to mimic the population in the Massachusetts Commonwealth Choice program. In the MEPS data, older consumers have higher medical expenditures but also pay a higher percentage of those medical expenditures out of pocket. Therefore, as a measure of relative costs *to the insurer*, we form the ratio of insured costs of older groups to the insured costs of the average insured costs of 27 to 30 year-old consumers.[30]

We find that the ratio of insured expenditure for the oldest consumer group (55-64 year olds) relative to those 30 and under is 2.7, implying that insurers would be constrained by a modified community rating $\theta = 2$ even in the absence of price discrimination motives. Yet the cost ratios for slightly younger consumers (i.e. 50-54 or 45-49) is much lower (about 1.5). This suggests that price discrimination explains part of the pricing pattern in the data. Specifically, consider 45-49 year old consumers. Cost estimates indicate that these consumers cost only slightly more (20%) to the insurer than consumers 27-30, yet premiums are 40% higher. This is easily rationalized by differences in elasticities: consumers age 27-30 have an elasticity that is over twice the elasticity for the older group.

Increasing the age-band θ reduces the transfer that young consumers give to older consumers. The level of θ is subject to regulation, and varies: while Massachusetts has imposed a 2:1 age band, PPACA calls for a 3:1 age band. Table 5 additionally shows how transfers change with modified community rating rules. The PPACA regulation lowers prices for consumers under 40 by approximately 3% relative to the Massachusetts regulation, before any cost differences are taken into account. Another way of describing the impact of preference heterogeneity is to consider its impact on regulation.

Any benefit of providing younger consumers with a separate market with

[30] A limitation of this analysis is that it does not account for differential selection into the exchange: the consumers who lacked coverage in the employer-based market are not representative of the population. However, in the absence of better cost data, it provides a useful baseline.

different types of plans should be weighed against the premium-reducing impact these consumers would have in the broader market, both from lower medical expenditures and higher price sensitivity. This is relevant not only when considering pooling across ages, but pooling across incomes as well. In Massachusetts, the subsidized market is separated from the unsubsidized market, but this need not be the case. If the markets were combined, assuming the subsidized consumers are more price sensitive, premiums would be lower for all of the unsubsidized consumers, leading to a redistribution from the subsidized consumers to unsubsidized consumers.

The discussion of age-based pricing regulation assumes that age-bands will bind if θ is less than the ratio of costs in the data. However, in the presence of imperfect competition and preference heterogeneity, regulation will bind for ratios of costs that are much lower than θ. We note that even smaller cost differences lead to binding regulation: for example, if insurers coarsely price to just 3 age groups, a θ of 2 will bind even if the ratio of costs is only 1.5, and θ of 3 will bind when the ratio of costs is only 2.5. This indicates that, given the cost differences in the MEPS, the Massachusetts age bands of 2:1 will certainly bind, and the federal age bands of 3:1 are likely to bind.

Finally, Figure 4 plots premiums by age assuming the costs in the MEPS data, the preferences in the demand system, and no regulation, under both perfect and imperfect competition. First, we note that there is a dramatic increase in costs at age 55. However, differences in preferences amplify the differences in costs, leading to a much larger gap between prices under perfect and imperfect competition. Furthermore, we not that the prices under imperfect competition are more extreme than those we observe in the data, which is reasonable given that they are unconstrained by regulation in this simulation. Finally, margins on the oldest consumers can be quite large at around \$100 per month, or 20% of the purchase price. By contrast, the margins on the youngest consumers are quite slim.

6 Market Participation and Markups

Our results are likely to represent a lower bound on the potential impact of modified community rating on younger consumers, as we have assumed all consumers purchase insurance–i.e. that the mandate is effective. In the absence of a mandate, consumers may opt out of coverage. If in the face of higher prices, younger consumers simply opt out of the market completely, this will exacerbate the transfers from the younger consumers left in the market to older consumers. The model above can be expanded to allow consumers to opt out of the market; for simplicity, we consider the full pooling case. Denote the participation rate of the young consumers by c_L and their take-up elasticity by ε_L. Then μ_L represents the percent markup under unconstrained pricing and $\overline{\mu}$ the percent markup under full pooling. The participation rate can then be written as

$$c_L = 1 - \varepsilon_L \left(\mu_L - \overline{\mu}\right).$$

The participation rate can be defined similarly for the older consumers. The optimal pooled markup can then be expressed as

$$\overline{\mu} = \frac{\sigma c_L \mu_L + (1-\sigma)\, c_H \mu_H}{\sigma c_L + (1-\sigma)\, c_h}.$$

We simulate the effect of allowing for opting out in the absence of a mandate in Figure 5. Simulation 1 (solid line), assumes full market participation among the older group. Using intermediate values from the previous simulations, we use a markup of 15% for the younger consumers and a percent markup of 35% for the older consumers. Estimates of health insurance take-up elasticities in the literature vary substantially, from near zero to -2 (Washington and Gruber 2005, Cutler and Reber 2002). However, as noted earlier in the paper, elasticities from the employer-sponsored insurance may not correspond well to this new environment. Therefore, we simulate the optimal markup under take-up elasticities ranging from zero (full compliance) to -5. We use optimal markups for each group as estimated above, as well as the empirical population shares. The relationship is nonlinear, but for a take-up elasticity of -5, no

younger consumers participate in the market and the optimal pooled markup is equal to the optimal markup for older consumers. Therefore, we can get a so-called death spiral from differences in preferences alone. Furthermore, to the extent that elasticities are correlated with underlying costs, imperfect competition can amplify the potential for a death spiral. In either case, our estimates of transfers are certainly a lower bound given the potential for less than full market participation.

Our results are more general and apply to any situation in which ε_L is greater in absolute magnitude than ε_H. Simulation 2 (dashed lines) represent a simulation in which both older and younger consumers are allowed to have nonzero take-up elasticities; however, the elasticity of the younger consumers is constrained to be exactly twice that of the younger consumers. So long as the older consumers are less responsive to price than the younger consumers, the death spiral pattern holds.[31] Our results emphasize the heterogeneity in consumer preferences for insurance that has been noted in the literature, and connects this idea to the response of insurers facing regulation in a new market, a health insurance exchange. Modified community rating rules have a large impact on this market even in the presence of a mandate. However, If the mandate were dropped or were not effective, heterogeneity in preferences alone can lead to a death spiral effect in which all price sensitive consumers exit the market.

7 Conclusion

This paper has analyzed consumer behavior using a novel data set in a health insurance exchange that servers as a model for national health reform. Our analysis of choices speaks to a number of policy and conceptual issues. First, calculating the insurance spending elasticity depends critically on the choice set available to consumers. Therefore, we use a discrete choice approach to

[31] To the extent that elasticities are correlated with underlying costs, the effects of selection are likely to be similar. In addition, the general effect of nonzero take-up elasticities for older consumers is to increase the numbers of older consumers in the market, as full pooling represents a price decrease for older consumers.

measure consumers' price sensitivity, and find broad evidence of consumer heterogeneity in preferences. A consumer in the 75th percentile of the price sensitivity distribution is four times more price sensitive than a counterpart in the 25th percentile.

Modified community rating and age-based pricing in the Commonwealth Choice program are crucially important regulations. Priceable variation in consumer preferences gives insurers an additional motive, due to price discrimination, to increase premiums to older consumers. Consistent with findings in the insurance literature, we find that consumers have very heterogeneous preferences. We extend these findings to argue that imperfect competition may amplify variation in insurance prices if inelastic consumers also tend to be high cost consumers. Therefore, regulators should be especially cautious in defining risk pools, as heterogeneous preferences for insurance can exacerbate differences in premiums among different groups of consumers.

Beyond identifying crucial regulatory features in HIEs, we also demonstrate *why* these regulations are important. Exchange designers must not only consider the nature of consumer demand, but also strategic insurer pricing in the face of both consumer demand and the regulatory regime. Understanding behavior on the exchanges is crucial to designing them well. Our results from Massachusetts provide an early look at HIE, and a foundation for researchers who study health exchanges and policymakers who design such exchanges.

References

Abaluck, J. and J. Gruber. 2011. Heterogeneity in Choice Inconsistencies among the Elderly: Evidence from Prescription Drug Plan Choice. The American Economic Review 101 (3): 377–381.

Blumberg, L.J., M. Buettgens, and B. Garrett. 2009. Age Rating Under Comprehensive Health Care Reform: Implications for Coverage, Costs, and Household Financial Burdens. Urban Institute Timely Analysis of Immediate Health Policy Issues. Available at: http://www. urban. org/UploadedPDF/411970_age _rating. pdf.

Buchmueller, T.C. and J. DiNardo. 1999. Did Community Rating Induce an Adverse Selection Death Spiral? Evidence from New York, Pennsylvania and Connecticut.

Bundorf, K., J. Levin, and N. Mahoney. 2008. Pricing, Matching and Efficiency in Health Plan Choice. mimeograph, Stanford University.

Carlin, C. and R. Town. 2007. Adverse selection, welfare and optimal pricing of employer-sponsored health plans. U. Minnesota Working Paper.

Chu, C.S., P. Leslie, and A. Sorensen. 2011. Bundle-Size Pricing as an Approximation to Mixed Bundling. The American Economic Review 101 (1): 263–303.

Cutler, D.M. and S.J. Reber. 1998. Paying for Health Insurance: The Trade-Off between Competition and Adverse Selection. Quarterly Journal of Economics 113 (2): 433–466.

Dafny, L., M. Duggan, and S. Ramanarayanan. 2009. Paying a premium on your premium? Consolidation in the US health insurance industry. National Bureau of Economic Research.

Dafny, L., K. Ho, and M. Varela. 2010. Let them have choice: Gains from shifting away from employer-sponsored health insurance and toward an individual exchange. National Bureau of Economic Research.

Duggan, M., P. Healy, and F.S. Morton. 2008. Providing prescription drug coverage to the elderly: America's experiment with Medicare Part D. The Journal of Economic Perspectives 22 (4): 69–92.

Einav, L., A. Finkelstein, and M.R. Cullen. 2010. Estimating Welfare in Insurance Markets Using Variation in Prices. Quarterly Journal of Economics 125 (3): 877–921.

Ericson, K.M.M. 2010. Market Design when Firms Interact with Inertial Consumers: Evidence from Medicare Part D.

Finkelstein, A. 2004. Minimum standards, insurance regulation and adverse selection: evidence from the Medigap market. Journal of Public Economics 88 (12): 2515–2547.

Finkelstein, A., J. Poterba, and C. Rothschild. 2009. Redistribution by insurance market regulation: Analyzing a ban on gender-based retirement annuities. Journal of financial economics 91 (1): 38–58.

Geruso, M. 2011. Community Rating in Employer Health Insurance: Inefficiencies Beyond Adverse Selection.

Gruber, J. 2011. Massachusetts points the way to successful health care reform. Journal of Policy Analysis and Management.

Gruber, J. and E. Washington. 2005. Subsidies to employee health insurance premiums and the health insurance market. Journal of Health Economics 24 (2): 253–276.

Herring, B. and M.V. Pauly. 2006. Incentive-compatible guaranteed renewable health insurance premiums. Journal of Health Economics 25 (3): 395–417.

Ilayperuma Simon, K. 2005. Adverse selection in health insurance markets? Evidence from state small-group health insurance reforms. Journal of Public Economics 89 (9): 1865–1877.

Kaufman, D.W., J.P. Kelly, L. Rosenberg, T.E. Anderson, and A.A. Mitchell. 2002. Recent patterns of medication use in the ambulatory adult population of the United States. JAMA: The Journal of the American Medical Association 287 (3): 337.

Ketcham, J.D., C. Lucarelli, E.J. Miravete, and M.C. Roebuck. 2010. Sinking, swimming, or learning to swim in Medicare Part D. American Economic Review.

Kolstad, J.T. and A.E. Kowalski. 2010. The impact of an individual health insurance mandate on hospital and preventive care: Evidence from Massachusetts. National Bureau of Economic Research.

Krueger, A.B. and I. Kuziemko. 2011. The Demand for Health Insurance among Uninsured Americans: Results of a Survey Experiment and Implications for Policy.

Long, S. and L. Phadera. 2009. Estimates of health insurance coverage in Massachusetts from the 2009 Massachusetts Health Insurance Survey. Commonwealth of Massachusetts, Division of Health Care Finance and Policy.

Lustig, J. 2010. The welfare effects of adverse selection in privatized Medicare. Department of Economics, Boston University.

Roe, C.M., A.M. McNamara, and B.R. Motheral. 2002. Gender-and age-related prescription drug use patterns. The Annals of pharmacotherapy 36 (1): 30.

Sethi-Iyengar, S., G. Huberman, and W. Jiang. 2004. How much choice is too much? Contributions to 401 (k) retirement plans. Pension design and structure: New lessons from behavioral finance, pp. 83–95.

Starc, A. 2010. Insurer Pricing and Consumer Welfare: Evidence from Medigap.

Strombom, B.A., T.C. Buchmueller, and P.J. Feldstein. 2002. Switching costs, price sensitivity and health plan choice. Journal of Health economics 21 (1): 89–116.

Zuckerman, S. and S. Rajan. 1999. An alternative approach to measuring the effects of insurance market reforms. Inquiry: a journal of medical care organization, provision and financing 36 (1): 44.

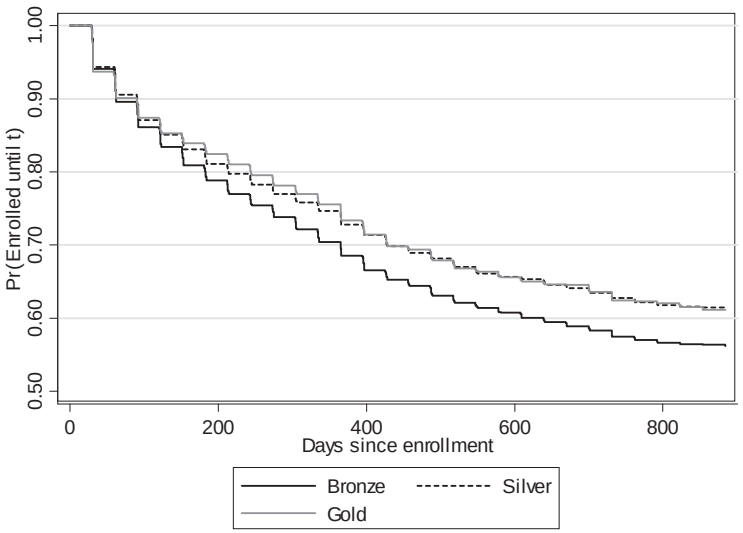


Figure 1: Length of Time Enrolled in the Connector by Tier of Plan. Notes: Kaplan-Meier survival estimate. Sample: all individuals, July 2007 to Dec 2009.

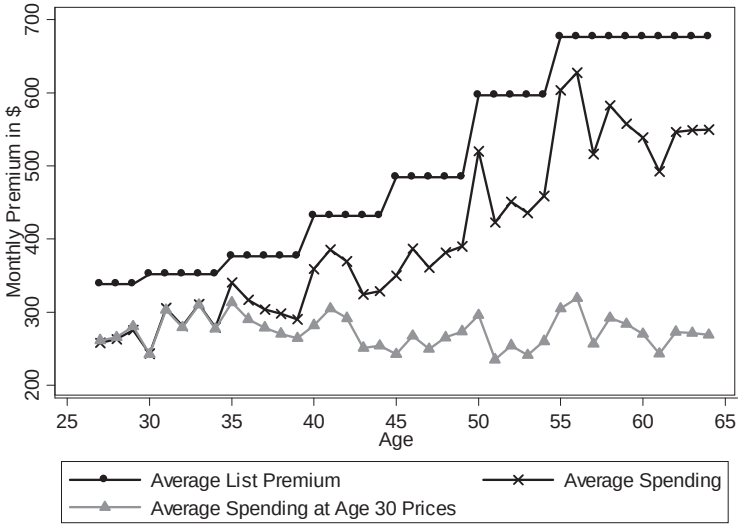


Figure 2: Average Prices and Spending, By Age. Notes: Data: Nov. and Dec. 2009. Average list premium is the plan-weighted average using the Nov. plan-zipcode price. Average spending is person-weighted. Average spending at age 30 prices uses actual choices but prices for a 30-year old in Nov. in that zipcode.

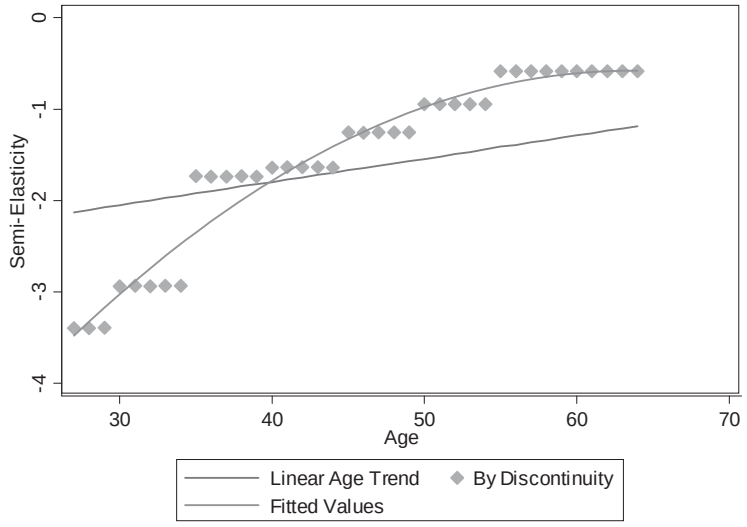


Figure 3: Semi-Elasticities By Age. Notes: Semi-elasticity describes behavioral response (reduction in market share) to a $100 increase in monthly premium. Linear age trend is plotted using results from Column 3 of Panel A of Table 3. Discontinuity results plot the average semi-elasticity obtained from Panel B of Table 3. The fitted values fit a quadratric trend to these estimates.

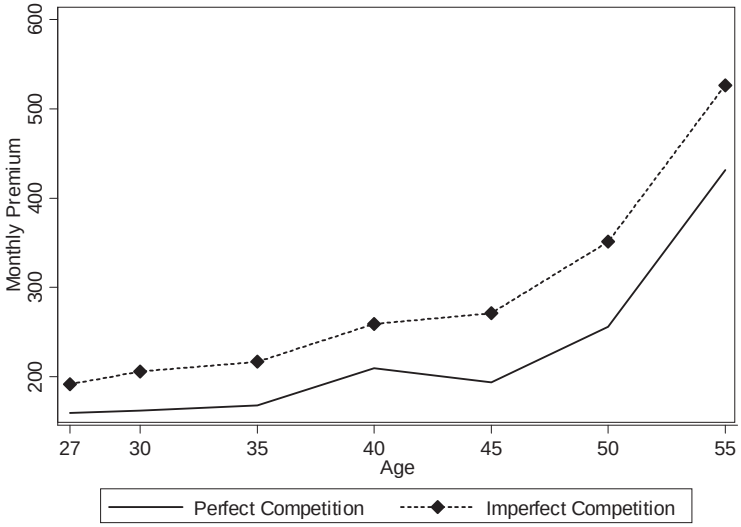


Figure 4: Simulated Monthly Premiums With Unconstrained Pricing. Notes: Assumes no age-based pricing regulation. Under perfect competition, we assume insurers charge at cost, with costs taken from the 2008 MEPS (see Table 6). Under imperfect competition, we assume firms charge the optimal markup based on age-specific price sensitivity calculated in Table 3 Panel B.

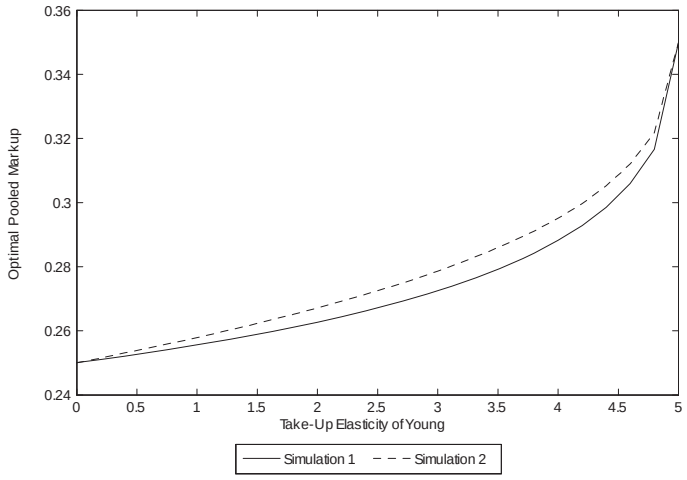


Figure 5: Optimal Markups in the Absence of a Mandate. Notes: Based on Simulation 1 and 2, described in the text.

Table 1: Demographics of the Connector

	Full Sample	Nov-Dec 2009
	Demographics	
Age	35.93	36.15
% Female	0.4524	0.4925
# of Lives Covered	1.311	1.291
Premium Paid (Monthly)	349.77	375.3
	Tiers	
Bronze	39.65	37.87
Bronze Plus	1.8	3.09
Silver	13.24	14.01
Silver Plus	2.86	2.36
Silver Select	5.16	5.51
Gold	7.42	7.01
Young Adult	29.87	30.15
	Insurers	
Blue Cross Blue Shield of Massachusetts	31.47	30.55
Fallon Community Health Plan	15.38	17.94
Harvard Pilgrim Health Care	22.7	21.19
Health New England	2.76	2.84
Neighborhood Health Plan	20.51	20.33
Tufts Health Plan	7.18	7.15

Note: Numbers represent simple averages from the raw data.

Table 2: Price and Spending Response to Age Discontinuities

	ln(price index)	ln(premiums paid)
Indicators:		
Above 30	0.0224***	-0.0438
	(0.00172)	(0.0323)
Above 35	0.0790***	0.0442
	(0.00199)	(0.0411)
Above 40	0.150***	0.147***
	(0.00217)	(0.0447)
Above 45	0.106***	0.0138
	(0.00189)	(0.0440)
Above 50	0.204***	0.207***
	(0.00201)	(0.0502)
Above 55	0.128***	0.192***
	(0.00232)	(0.0462)
Linear Age Spline	Yes	Yes
Basic Controls	Yes	Yes
N Persons	2,616	2,616
R^2	0.998	0.572

IV-Stage 1 from Column 1

	ln(premiums paid)
ln(price index)	0.962
	(0.176)
Linear Age Spline	Yes
Basic Controls	Yes
R^2	0.569

Sample: July-Dec 2009. Note: Heteroskedasticity robust standard errors in parentheses. Age spline consists of piecewise linear age controls within each age group. Controls include indicators for month of enrollment, indicators for geographic market, and gender. IV results from two-stage least squares. *** p<0.01, ** p<0.05, * p<0.1.

Table 3: Price Sensitivity by Age in Conditional Logit Model

	Panel A: Basic Conditional Logits (All Ages)		
	(1)	(2)	(3)
Premium	-0.357***	-2.018***	-2.266***
(in $100s)	(0.122)	(0.306)	(0.369)
Premium*age		0.0298***	0.0267**
		(0.00488)	(0.0114)
Fixed Effects	Plan	Plan	Plan, Plan*Age
N Person*Plan	20,838	20,838	20,838

	Panel B: Conditional Logits by Age Group					
	27-34	30-39	36-44	40-49	46-54	50+
Premium	-3.574***	-2.611***	-2.354***	-2.271***	-1.512***	-1.234***
(in $100s)	(0.533)	(0.560)	(0.606)	(0.508)	(0.572)	(0.316)
N Person*Plan	8,512	5,396	4,380	4,459	3,745	5,628
Plan and Tier*Age^2	Yes	Yes	Yes	Yes	Yes	Yes

	Panel C: Conditional Logits For Counterfactual Exercise	
	Under Age 45	Age 45+
Premium	-2.747***	-0.752***
(in $100s)	(0.382)	(0.266)
Plan and Tier*Age^2	Yes	Yes
N Person*Plan	12,892	7,946

Note: Heteroskedasticity robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Panels B and C include plan fixed effects, and tier effects interacted with age trends (both linear and quadratic terms).

Table 4: Age-based Price Sensitivity in Additional Specifications

	(1)	(2)	(3)	(4)
	CL	CL	NL	ML
Premium in Hundreds of $	-3.211***	-3.207***	-1.398***	
	(0.416)	(0.456)	(0.165)	
...*1(30-34)	0.602***	0.440*	0.138	0.475***
	(0.181)	(0.237)	(0.119)	(0.180)
...*1(35-39)	0.779***	0.461	0.195	0.585***
	(0.255)	(0.378)	(0.130)	(0.202)
...*1(40-44)	0.938***	0.572	0.326***	0.609***
	(0.328)	(0.461)	(0.121)	(0.210)
...*1(45-49)	1.084***	0.780	0.218*	0.470**
	(0.390)	(0.492)	(0.121)	(0.216)
...*1(50-54)	1.471***	1.305**	0.398***	0.736***
	(0.450)	(0.515)	(0.128)	(0.242)
...*1(55+)	1.892***	1.855***	0.707***	1.197***
	(0.500)	(0.519)	(0.0994)	(0.211)
Bronze Dissimilarity Parameter			0.531***	
			(0.0777)	
Silver Dissimilarity Parameter			0.608***	
			(0.0933)	
Gold Dissimilarity Parameter			0.977***	
			(0.208)	
Mean Premium Parameter				1.019***
				(0.0945)
S.D. Premium Parameter				0.345***
				(0.0400)
Fixed Effects	Plan	Plan		Tier
	Plan*Age	Plan*Age	Tier*Age	Tier*Age
		Plan*Age^2	Carrier	Carrier
N Person*Plan	20,838	20,838	20,838	20,838

Note: Heteroskedasticity robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Columns 1 and 2 contain conditional logit specifications. Column 3 estimates and nested logit and reports dissimilarity parameters, where a dissimilarity parameter equal to one collapses to a conditional logit. Column 4 reports a mixed logit specification in which the price coefficient is allowed to take on a lognormal distribution. Additional mixed logit specifications by age confirm the general pattern in Table 3.

Table 5: Transfers From Under 45 Year Olds, as a Percent of Premiums

	Firms Price to:	
	2 Age Groups	3 Age Groups
Full Pooling	6.01%	7.55%
2:1 Age Bands	4.80%	5.98%
3:1 Age Bands	1.43%	3.20%

Notes: Transfers are calculated as the differences in optimal firm markups under each regulatory policy. Transfers are calculated assuming constant costs across ages.

Table 6: Comparison of Costs Across Age Groups

	Population (in thousands)	Avg. Annual Expense ($ per capita)	Percent Paid... Out of pocket	Percent Paid... by Private Insurance	Avg. Insurer Cost c_a	Ratio: c_a/c_{27}
Total	98968	$ 3,992	18.5	76.5	$ 3,054	
Age group:						
27-30	7226	$ 2,401	17.4	79.7	$ 1,914	1.00
31-34	11715	$ 2,509	18.3	77.5	$ 1,945	1.02
35-39	12866	$ 2,723	23.3	73.9	$ 2,012	1.05
40-44	13863	$ 3,279	18.7	76.6	$ 2,512	1.31
45-49	13652	$ 3,241	19.6	71.6	$ 2,320	1.21
50-54	15092	$ 4,046	18.9	75.9	$ 3,071	1.60
55-64	24554	$ 6,627	17.2	78.1	$ 5,175	2.70

Note: Data taken from 2008 MEPS, with authors' calculations. Sample selection: people age 27-64 with middle or high incomes with any private insurance in the New England states. Avg. insurer cost is mean priviate insurer expenditure for this sample.