Experimental Design as Market Design: Billions of Dollars Worth of Treatment Assignments*

Yusuke Narita†

October 17, 2017

Abstract

Randomized Controlled Trials (RCT) enroll hundreds of millions of people and involve many human lives. In this paper, I propose a design of RCT with high-stakes treatment. Unlike conventional RCT, my design respects subject welfare; it optimally randomly assigns each treatment to subjects predicted to experience better treatment effects, or to subjects with stronger preferences for the treatment. For preference elicitation, my design is also almost incentive compatible. Finally, this design unbiasedly estimates any causal effect estimable with standard RCT. To quantify these properties, I apply my proposal to a water cleaning experiment in Kenya (Kremer et al., 2011). Compared to usual RCT, my design substantially improves subjects' well-being while reaching similar treatment effect estimates with similar precision.

Keywords: Clinical Trial, Medical Ethics, Field and Social Experiment, Mechanism Design, Causal Inference, Statistical Decision, Development Economics, Water Source Protection, Discrete Choice

Work in Progress
Comments Welcome

*I am grateful to Dean Karlan for a conversation that inspired this project; Joseph Moon for industrial and institutional input; Sylvain Chassang, Costas Meghir, and Joseph Shapiro for extensive comments; seminar participants at Yale, Hitotsubashi, European Summer Symposium in Economic Theory (ESSET) on "An Economic Perspective on the Design and Analysis of Experiments," CyberAgent, Kyoto, International Conference on Experimental Methods in Economics. I received helpful research assistance from Shreyas Ravishankar, Soumitra Shukla, Jaehee Song, Mizuhiro Suzuki, Devansh Tandon, Lydia Wickard, and especially Qiwei He, Zaiwei Liu, Vincent Tanutama, and Kohei Yata.

†Yale University, Department of Economics and Cowles Foundation. Email: yusuke.narita@yale.edu

1 Introduction

Now is the golden age of Randomized Controlled Trials (RCT; equivalently, randomized experiments or A/B tests). Having originally started as a safety and efficacy test of farming and medical treatments (Gaw, 2009), RCT has grown to become the gold standard of evidence-based decisions and evaluations. RCT is widespread in business and politics (Siroker and Koomen, 2013), as well as public policy (Gueron and Rolston, 2013), economics (Banerjee and Duflo, 2012), other social sciences (Gerber and Green, 2012), and engineering.

RCTs are high-stakes. Firstly, a large number of individuals participate in RCTs. For example, I find that above 360 million patients in total participate in clinical trials registered in WHO's International Clinical Trials Registry Platform (ICTRP) during 2007-2017.[1] As for social and economic RCTs, over 22 million individuals participate in experiments registered in American Economic Association's registry for the last decade.

To such a big subject pool, RCT sometimes randomizes high-stakes and even life-or-death treatment. For instance, in a glioblastoma therapy trial, the five-year death rate of glioblastoma patients is 97% in the control group but only 88% in the treatment group (Stupp et al., 2009). This means that in expectation, the lives of 9% of its 573 participants depend on who receive treatments. Social and economic RCTs also sometimes randomize critical treatment such as basic income[2], health insurance (Baicker et al., 2013), high wage job offers (Dal Bó et al., 2013), HIV testing (Angelucci and Bennett, 2017), and cash transfers (Haushofer and Shapiro, 2016). As a consequence of the high-stakes nature, some RCTs resulted in media controversies and lawsuits by participants (for instance, *Gelsinger v. University of Pennsylvania* and *Grimes v. Kennedy-Krieger Institute*).[3]

RCT thus determines the fate of numerous people. Physician and prior editor-in-chief of the *New England Journal of Medicine*, Marcia Angell, noted the resulting ethical dilemma:

> *How can a physician committed to doing what he thinks is best for each patient tell a woman with breast cancer that he is choosing her treatment by something like a coin toss? How can he give up the option to make changes in treatment according to the patient's responses?* ("Patients' Preferences in Randomized Clinical

[1] Related to the subtitle of this paper, clinical trials are a costly investment. For the US, for example, the average trial cost is believed to be at least thousands of dollars per subject (Morgan et al., 2011). Multiplying a few thousands of dollars by the number of subjects in the US (still over 300 millions in total), the total investments into clinical trials may amount to 1 trillion dollars for the last decade just in the US.

[2] "8 basic income experiments to watch out for in 2017," at `http://www.businessinsider.com/basic-income-experiments-in-2017-2017-1/#finland-2`, retrieved October 2017.

[3] *Gelsinger v. University of Pennsylvania* was about a gene-therapy clinical trial while *Grimes v. Kennedy-Krieger Institute* about a social experiment that randomly assigned lead reduction methods to housings. For details, see `https://www.sskrplaw.com/publications.html` and `http://www.courts.state.md.us/opinions/coa/2001/128a00.pdf`, accessed in October 2017.

Trials")

The welfare impact of RCTs motivates me to study how to care about subject well-being. I formulate experimental design as the problem of computing treatment assignment probabilities based on data about the predicted treatment effect of each treatment on each subject and each subject's willingness-to-pay (WTP) for each treatment. Predicted effects and WTP may be arbitrarily correlated and come from prior experimental, observational, or self-report data.

I propose a data-driven experimental design which I call *Experiment-as-Market (EXaM).*[4] I call my EXaM design Experiment-as-Market because EXaM uses economic ideas on market to solve the statistical and empirical problem of experimental design. EXaM randomly assigns treatments to subjects via an imaginary centralized market (inspired by the long-standing idea of competitive market equilibrium from equal incomes by Friedman (1962); Varian (1974); Hylland and Zeckhauser (1979); Budish et al. (2013); He et al. (2017)). EXaM endows each subject with a common artificial budget and lets her use the budget to purchase a most preferred bundle of treatment assignment probabilities given their prices. The prices are discriminated so that each treatment is cheaper for subjects predicted to experience better effects of the treatment. EXaM computes its treatment assignment probabilities as what subjects demand at market clearing prices, where subjects' aggregate demand for each treatment is balanced with its capacity (assumed to be exogenously given).

This virtual-market construction gives EXaM nice welfare and incentive properties. Unlike standard RCT, EXaM has a Pareto optimality property that no other design makes every subject better-off in terms of expected predicted effects of and WTP for assigned treatment. EXaM also allows the experimenter to elicit WTP in an almost incentive compatible way. That is, when the experimenter asks subjects to self-report their WTP to be used by EXaM, every subject's optimal choice is to report her true WTP, at least for large experiments.[5]

EXaM also allows the experimenter to unbiasedly estimate the same treatment effects as standard RCT does (in a wide class of treatment effect parameters). Since EXaM gives everybody the same budget, if subjects share the same predicted effects and WTP, the subjects purchase the same distribution of treatment assignment. In other words, EXaM produces treatment assignment that is independent (unconfounded) from potential outcomes condi-

[4]For running EXaM, I assume the experimenter has data on the set of subjects and treatments, treatment capacities, each subject's WTP for each treatment, and the predicted effect of each treatment for each subject. EXaM is executable, however, even without WTP and predicted effects (perhaps when WTP and predicted effects are unknown or irrelevant to the experimenter). When the experimenter uses neither WTP nor predicted effects, EXaM reduces to standard RCT. Therefore, EXaM nests standard RCT.

[5]The analysis of EXaM's incentive property owes much to studies on the incentive compatibility of competitive equilibria (Roberts and Postlewaite, 1976; Jackson, 1992; Azevedo and Budish, 2017; He et al., 2017).

tional on observable predicted effects and WTP. This property puts EXaM in the context of causal inference with stratified experiments and selection-on-observables, as summarized by Imbens and Rubin (2015).

Specifically, the conditionally independent treatment assignment in EXaM allows the experimenter to unbiasedly estimate the conditional average treatment effects conditional on observables. The estimable conditional effects turn out to contain Marginal Treatment Effects (Björklund and Moffitt, 1987; Heckman and Vytlacil, 2005) as a subset. By integrating such marginal or conditional effects, EXaM can also unbiasedly estimate the (unconditional) average treatment effect, the single most important causal effect estimable with RCT. Perhaps more importantly, EXaM's unbiased average treatment effect estimation may have a smaller standard error than RCT's.[6]

I compare EXaM not only with RCT but also with more elaborate designs that pay attention to WTP or predicted effects. Such sophisticated designs include Randomized Consent Trials (Zelen, 1979; Angrist and Imbens, 1991), Selective Trials (Chassang et al., 2012), and Multi-Armed Bandit algorithms (White, 2012). Compared to these existing designs, EXaM integrates the WTP and predicted effect considerations into a unified design in an optimal and incentive compatible way.

Finally, as a proof of concept, I empirically apply EXaM to data from a water cleaning experiment in Kenya (Kremer et al., 2011). Compared to RCT, EXaM turns out to substantially improve participating households' welfare, which is measured by predicted child diarrhea reduction by water cleaning and revealed WTP for it. Data from EXaM also allows me to get similar treatment effect estimates with similar standard errors as RCT does. Along the way, I develop a computer program to implement EXaM with little computational cost.

Taken together, EXaM sheds light on a way economic thinking can "*facilitate the advancement and use of complex adaptive (...) and other novel clinical trial designs,*" one of the federal Food and Drug Administration (FDA)'s performance goals for 2018-2022.[7] More concretely, my analysis shows the value of using subjects' WTP for treatments. The use of WTP in EXaM complements existing uses in Ashraf et al. (2006); Cohen and Dupas (2010); Ashraf et al. (2010); Chassang et al. (2012); Devoto et al. (2012); Dupas (2014). EXaM combines the preference consideration with another idea of respecting predicted effects. The use of predicted effects is becoming established in medicine (Food and Drug Administration,

[6]These informational values materialize regardless of whether the experimenter correctly predicts treatment effects and WTP. This experimental value of EXaM and competitive equilibrium from equal incomes echoes Abdulkadiroğlu et al. (2017) and Narita (2016), who highlight the informational values of a different sort of mechanism design (centralized school choice with lotteries).

[7]See `https://www.fda.gov/downloads/forindustry/userfees/prescriptiondruguserfee/ucm511438.pdf`, retrieved in October 2017

2010) and business (White, 2012) but just emerging in social sciences (Manski, 2008).

The next section motivates my agenda with facts about the impact of RCT on participant welfare. Following this context, Section 3 develops the EXaM experimental design while Section 4 shows its welfare and incentive properties. Section 5 turns to the experimental information embedded in EXaM and explains how to use data from EXaM for causal inference. After comparing EXaM with existing designs in Section 6, Section 7 presents an empirical application. Finally, Section 9 summarizes my findings, discusses their limitations, and outlines future directions. Proofs are in Appendix A.1.2.

2 Why Subject Welfare?

I study experimental design with an emphasis on subject welfare. Why should I study subject well-being? This section demonstrates facts that show that treatment allocation in RCTs impacts the lives of numerous individuals.

Normative Considerations

First of all, RCTs involve a large number of subjects. To see it, I assemble data on clinical trials registered in the WHO International Clinical Trials Registry Platform (ICTRP).[8] ICTRP is the largest international clinical trial registry and subsumes domestic platforms like ClinicalTrials.gov for the US.[9] Table 1 shows that the number of registered trials amounts to about 290 thousands from 2007 to mid-2017. The sum of their sample sizes is over 360 millions for the same period.[10]

It is important to note that the figures in Table 1 are likely to underestimate the total scale of the clinical trial landscape. Many countries (such as Australia and Japan) do not legally require clinical trials to register (as of October 2017). Even when trials are required to register, the expected fine for failing to do so is often negligible compared to the total trial cost.[11] As a consequence of these regulatory loopholes, there is likely a "dark pool" of clinical trials never reflected in any public database like ICTRP (Goldacre, 2014).[12]

[8] `http://www.who.int/ictrp/en/`, retrieved in October 2017

[9] `https://clinicaltrials.gov`, retrieved in October 2017

[10] More detailed statistics are in Appendix Tables A.1-A.3. I also find the sum of sample sizes of registered *economic RCTs* amounts to above 23 millions for the last decade (Appendix Table A.4).

[11] See *Stat News*' article, "Failure to report: A STAT investigation of clinical trials reporting," at `https://www.statnews.com/2015/12/13/clinical-trials-investigation/`, retrieved October 2017.

[12] Consistent with this hypothesis, as legal and institutional pressures for trial registration mount, the annual numbers of registered trials and subjects are rapidly growing (about 14 millions in 2007 vs. 72 millions in 2016 for the number of subjects; see Figure 1.b). This also means that these figures will likely be larger in the next decade. As a conservative scenario, assume the annual sum of sample sizes will stay at the 2016 level (72 millions). The total sample size for the next decade will be 720 millions ($= 10 \times 72$ millions)

For such a large subject population, RCTs frequently randomize high-stakes treatment. The high-stakes and occasionally life-threatening nature of many RCTs is highlighted by examples in Table 2. In the first clinical trial (row i in Panel (a)), for example, a cholesterol lowering drug treatment was found to lower the 5 year death rate by about 30% relative to the baseline death rate in the control group. Other clinical trials in Table 2 Panel (a) also report significant impacts on survival and other crucial outcomes.[13]

Even social and economic RCTs randomized treatment such as cash transfers, health insurance, HIV testing, and police patrol, as can be seen in Table 2 Panel (b). As expected, these treatments are often found to have profound treatment effects. In addition to these published examples, several RCTs of basic income are also ongoing or announced (recall footnote 2).

Practical Considerations

Practical considerations also motivate a care for subject welfare. The successful implementation of any RCT depends on subject choices and behavior such as whether subjects participate in the RCT; whether subjects take up and use the assigned treatment; and whether subjects stay in contact in a follow-up period. The RCT produces useful information only if participants are active enough in each step. This prerequisite is hard to achieve, however. Many RCTs suffer from subject indifference or fear in the form of non-participation, non-compliance, and dropouts before, during, and after the experiment (Friedman et al. (1998) chapters 10 and 14 and Duflo et al. (2007) sections 4.3 and 6.4).

A welfare-conscious experimental design could alleviate non-participation, non-compliance, and dropouts.[14] In fact, King et al. (2005) provide a clinical trial meta-analysis suggesting that incorporating subject preferences makes subject recruitment easier. Chan and Hamilton (2006) suggest that better-off subjects experiencing better treatment effects are less likely to drop out. In a range of econometric and theoretical models (Heckman and Vytlacil, 2005; Chan and Hamilton, 2006), welfare-enhancing treatment assignment is predicted to facilitate compliance with treatment assignment.[15]

[13]The medical ethics literature reviews other examples (Shamoo and Resnik (2009) chapters 12 and 13 and Freeman et al. (2001)).

[14]In an effort to minimize attrition and maximize the treatment take-up rate, many field experiments start with an expression-of-interest survey before randomization and recruit only survey respondents who express strong interest (Duflo et al., 2007). This recruitment practice causes external validity concerns. These concerns may also be alleviated by replacing the experimenter's discretionary selective recruitment with an experimental design respecting subject welfare in a rule-based way, as I do in this paper.

[15]In addition, more ethical experimental designs would ease collaboration with partner governments and companies that may have an ethical and reputational concern with involvement in RCTs.

3 Experiment-as-Market (EXaM)

3.1 Framework

The normative and practical importance of subject well-being prompts me to design an experiment that balances experimental information with subject welfare. An *experimental design problem* consists of:

- Experimental *subjects* $i_1, ..., i_n$.
- Experimental *treatments* $t_0, t_1, ..., t_m$ where t_0 is a placebo or control.
- Each treatment t's *capacity* or *supply* $c_t \in \mathbb{N}$ with $\Sigma_{t=t_0}^{t_m} c_t = n$.
- Each subject i's *preference or WTP* $w_{it} \in \mathbb{R}$ for treatment t where $w_{it} \geq w_{it'}$ means subject i weakly prefers treatment t over t'. Let $w_i \equiv (w_{it})_t$.
- Each treatment t's *predicted treatment effect* $e_{ti} \in \mathbb{R}$ for subject i where $e_{ti} \geq e_{t'i}$ means treatment t is predicted to have a weakly better effect than t' for subject i.[16]

I normalize e_{ti} and w_{it} by assuming $e_{t_0 i} = w_{it_0} = 0$ for every subject i. e_{ti} and w_{it} are therefore predicted effects of t and WTP for t, respectively, relative to the control t_0. This normalization is without loss of generality because only differences in WTP and predicted effects matter for subject welfare from inside treatments $t_0, t_1, ..., t_m$. Every experimental design discussed in this paper produces the same treatment assignment probabilities with and without the normalization.

A few remarks are in order. First of all, where do WTP and predicted effects come from? As for WTP w_{it}, there are a few possible sources. The experimenter may estimate WTP from data on treatment choices by subjects, as I do with a discrete choice model in my empirical application in Section 7.[17] Alternatively, the experimenter may ask each subject i to self-report WTP w_i, as proposed by Zelen (1979) and Chassang et al. (2012).[18]

On predicted effects e_{ti}, it is best to estimate them from prior experimental or observational data. The most reliable data source is a prior RCT of a similar treatment. Such sequential RCTs with the same treatment are common in medicine (Friedman et al., 1998)

[16] Here I assume WTP and predicted effects are fixed and with cardinal meaning. See Section 8 for what to do when WTP and predicted effects are uncertain or ordinal.

[17] Similar demand estimation but for different purposes can be found in Ashraf et al. (2006); Cohen and Dupas (2010); Ashraf et al. (2010); Kremer et al. (2011); Devoto et al. (2012); Dupas (2014).

[18] This self-report method raises the question of incentive compatibility. I study incentive compatibility theoretically in Sections 4.2 and empirically in Section 7.3.

and business (Siroker and Koomen, 2013) and are growing in social sciences (Hahn et al., 2011). I illustrate the use of such a prior RCT in my empirical application in Section 7.[19]

Finally, predicted effects and WTP may be freely correlated. This is an important generality since evidence of selection or correlation between treatment effects and WTP is ample both in economics and medicine (Preference Collaborative Review Group, 2008; Swift and Callahan, 2009). To be consistent with the evidence, the above setup allows arbitrary selection.

3.2 Experimental Designs

An *experimental design* maps each experimental design problem into treatment assignment probabilities (p_{it}) satisfying the capacity constraint $\Sigma_i p_{it} \leq c_t$ for every treatment t. Here p_{it} is the probability that subject i is assigned to treatment t under the experimental design. The benchmark design is the standard Randomized Controlled Trial (RCT), formalized as follows.

Definition 1 (*Randomized Controlled Trial* a.k.a. *RCT*)**.** *Randomized Controlled Trial* is an experimental design that assigns each subject i to each treatment t with the impersonal treatment assignment probability $p_{it}^{RCT} \equiv c_t/n$.

I investigate welfare-enhancement with an alternative design, which I call Experiment-as-Market or EXaM in short.

Definition 2 (*Experiment-as-Market* a.k.a. *EXaM*)**.** (1) In a computer, distribute common artificial budget $b > 0$ to every subject. Find any price-discriminated competitive market equilibrium, i.e., any treatment assignment probabilities (p_{it}^*) and their prices π_{te} with the following properties:

- Effectiveness-discriminated treatment pricing: There exist $\alpha < 0$ and $\beta_t \in \mathbb{R}$ for each treatment t such that the price of a unit of probability of assignment to t for subjects with $e_{ti} = e \in \mathbb{R}$ is

$$\pi_{te} = \alpha e + \beta_t.$$

- Subject utility maximization subject to the budget constraint: For all subject i,

$$(p_{it}^*)_t \in \arg\max_{p_i \in P} \Sigma_t p_{it} w_{it} \text{ s.t. } \Sigma_t p_{it} \pi_{te_{ti}} \leq b,$$

[19] DellaVigna and Pope (2016) investigate an alternative approach of asking experts to forecast treatment effects.

where $p_i \equiv (p_{it})_{t=t_0,t_1,\ldots,t_m}$ and $P \equiv \{p_i \in \mathbb{R}^{m+1} | \Sigma_{t=t_0}^{t_m} p_{it} = 1\}$ is the set of feasible treatment assignment probability vectors (m-dimensional simplex) for each subject. $\pi_{te_{t_1 i}}$ is the price of a unit of the probability of assignment to treatment t for subject i. EXaM breaks ties by uniformly mixing utility-maximizing p_i's that solve the above problem with the minimum expenditure $\Sigma_t p_{it} \pi_{te_{ti}}$.

- Meeting capacity constraints: $\Sigma_i p_{it}^* \leq c_t$ for every treatment t.

(2) Compute

$$p_{it}^*(\epsilon) \equiv (1-q) p_{it}^* + q p_{it}^{RCT},$$

where $q \equiv \inf\{q' \in [0,1] | (1-q') p_{it}^* + q' p_{it}^{RCT} \in [\epsilon, 1-\epsilon] \text{ for all } i \text{ and } t\}$. Here $\epsilon \in [0, \bar{\epsilon})$ is a parameter fixed by the experimenter where $\bar{\epsilon} \equiv \min\{\min_t p_{it}^{RCT}, 1 - \max_t p_{it}^{RCT}\}$ is the largest possible value of ϵ.[20]

I name this experimental design Experiment-as-Market (EXaM) because EXaM randomly assigns treatments to subjects via a synthetic centralized market. This market builds up on the classic idea of competitive market equilibrium from equal incomes, owing much to the literature comprising Friedman (1962), Varian (1974), Hylland and Zeckhauser (1979), Budish et al. (2013), He et al. (2017) among others. EXaM can be seen as a generalization or variation of their ideas.

More specifically, in Step 1 of Definition 2, EXaM endows each subject with a common artificial budget. EXaM then lets each subject use the budget to purchase a most preferred bundle of treatment assignment probabilities taking their prices as given. The prices are partially personalized so that each treatment is cheaper for subjects predicted to benefit more from the treatment. EXaM computes its treatment assignment probabilities as those subjects purchase at market clearing prices, where subjects' total demand for each treatment is balanced with its given supply.[21] EXaM finally requires each subject to get each treatment with a probability strictly between 0 and 1, as done in Step 2. This requirement is important

[20] Why is $\bar{\epsilon}$ the largest possible value of ϵ? To see the reason, suppose $\epsilon > \min_t p_{it}^{RCT}$. Then, for any $t \in \arg\min_t p_{it}^{RCT}$, whenever $p_{it}^* \leq p_{it}^{RCT}$, I have

$$(1-q') p_{it}^* + q' p_{it}^{RCT} \notin [\epsilon, 1-\epsilon]$$

for any $q' \in [0,1]$. Similarly, if $\epsilon > 1 - \max_t p_{it}^{RCT}$, then for any $t \in \arg\max_t p_{it}^{RCT}$, whenever $p_{it}^* \geq p_{it}^{RCT}$,

$$(1-q') p_{it}^* + q' p_{it}^{RCT} \notin [\epsilon, 1-\epsilon]$$

for any $q' \in [0,1]$. Thus ϵ cannot exceed $\bar{\epsilon} \equiv \min\{\min_t p_{it}^{RCT}, 1 - \max_t p_{it}^{RCT}\}$.

[21] The first step of Definition 2 raises two questions, whether such an equilibrium exists and how to find such an equilibrium. After positively solving the first existence question in Proposition 2 below, I develop and implement a script to find an equilibrium in the empirical application in Section 7. See Budish et al. (2016) for a related algorithmic development on a different problem (MBA course allocation).

for EXaM to produce non-degenerate random assignment and unbiasedly estimate causal treatment effects.[22]

EXaM is an enrichment of RCT. To see this, note that EXaM allows the experimenter to turn off welfare considerations. For instance, if the experimenter does not know or care about predicted effects, she would let $e_{ti} = e_{tj}$ for all subjects i and j and treatment t. Similarly, let $w_{it} = w_{jt}$ if WTP is unknown or irrelevant. The following fact shows that EXaM is equivalent to RCT when the experimenter ignores both WTP and predicted effects.

Proposition 1 (EXaM nests RCT)**.** *Suppose that WTP and predicted effects are unknown or irrelevant so that $w_{it} = w_{jt}$ and $e_{ti} = e_{tj}$ for all subjects i and j and treatment t. Then EXaM reduces to RCT, i.e., for every $\epsilon \in [0, \bar{\epsilon})$, subject i, and treatment t, I have*

$$p^*_{it}(\epsilon) = p^{RCT}_{it}.$$

4 Welfare and Incentive: Theory of EXaM

4.1 Welfare

As opposed to the special case in Proposition 1, the experimenter is often concerned about WTP and predicted effects (as in studies reviewed in Section 2). In such cases, EXaM differs from RCT. Specifically, EXaM respects subject welfare by optimally assigning each treatment to subjects with higher WTP for the treatment, or to subjects predicted to benefit more from the treatment.

Proposition 2 (Existence and Welfare)**.** *EXaM $p^*_{it}(\epsilon)$ exists for any experimental design problem and any $\epsilon \in [0, \bar{\epsilon})$. Moreover, there is no other experimental design $(p_{it}) \in P^n$ with $p_{it} \in [\epsilon, 1-\epsilon]$ for all subject i and treatment t and the following better welfare property:*

$$\Sigma_t p_{it} w_{it} \geq \Sigma_t p^*_{it}(\epsilon) w_{it} \text{ and } \Sigma_t p_{it} e_{ti} \geq \Sigma_t p^*_{it}(\epsilon) e_{ti}$$

for all i with at least one strict inequality.[23]

[22] The definition leaves unspecified how to draw a final treatment assignment from $p^*_{it}(\epsilon)$. It is known to be always possible to draw a treatment assignment in a way consistent with $p^*_{it}(\epsilon)$ (Budish et al. (2013)'s Theorem 1, the generalized Birkhoff-von Neumann Theorem). For the moment, my analysis applies to any method to draw a treatment assignment. I impose more structures in Section 5 and implement an algorithm to draw an assignment in the empirical application in Section 7.

[23] Another version of the welfare result is also true: There is no other experimental design $(p_{it}) \in P^n$ with $p_{it} \in [\epsilon, 1-\epsilon]$ for all subject i and treatment t and the following better welfare property: $\Sigma_t p_{it} w_{it} \geq \Sigma_t p^*_{it}(\epsilon) w_{it}$ for all i and $\Sigma_i p_{it} e_{ti} \geq \Sigma_i p^*_{it}(\epsilon) e_{ti}$ for all t with at least one strict inequality.

Proposition 2 says that no other experimental design ex ante Pareto dominates EXaM in terms of WTP or predicted effects (while satisfying the random assignment condition). This ex ante Pareto optimality is known to imply ex post Pareto optimality and "ordinal" ex ante optimality (Bogomolnaia and Moulin, 2001).[24] In contrast, RCT fails to satisfy the welfare property as it ignores WTP and predicted effects. I empirically quantify the welfare gap between RCT and EXaM in Section 7.3.

Proposition 2 uses e_{ti} and w_{it} as two joint welfare measures, one outcome- or treatment-effect-based and one WTP-based. Each of them has an established role in economic welfare analysis. The medical literature more frequently studies treatment effects but also emphasizes that patients often have preferences for treatments (even conditional on treatment effects). This is especially the case for psychologically sensitive treatments like abortion methods (Henshaw et al., 1993) and depression treatments (Chilvers et al., 2001). In response to these intuitive findings, a US-government-endorsed movement tries to bridge the gap between evidence-based medicine and patient-preference-centered medicine (Food and Drug Administration, 2016). According to advocates, "*patient-centered care (...) promotes respect and patient autonomy; it is considered an end in itself, not merely a means to achieve other health outcomes*" (Epstein and Peters, 2009). My welfare criterion echoes this trend and accommodates both outcome- and preference-based approaches.

4.2 Incentive

So far I take WTP w_{it} as given and assume it to represent true WTP. In practice, the experimenter often needs to elicit the WTP information from subjects, raising an incentive compatibility concern. EXaM turns out to allow the experimenter to extract WTP in an almost incentive compatible way. My analysis of incentive compatibility builds up on the literature on incentive compatibility of competitive equilibria (Roberts and Postlewaite, 1976; Jackson, 1992; Azevedo and Budish, 2017; He et al., 2017)

Unfortunately, it is known that no experimental design satisfies the welfare property in Proposition 2 and exact incentive compatibility for general problems. This compels me to investigate approximate incentive compatibility in large experimental design problems. Consider any sequence of experimental design problems $(i_1, ..., i_n, t_0, t_1, ..., t_m, (c_t^n), (w_{it}^n), (e_{ti}^n))_{n\in\mathbb{N}}$ indexed by the number of subjects, n. The set of treatments $t_0, t_1, ..., t_m$ is fixed, but ev-

[24]Here ex post optimality means that no other (p_{it}) with the ϵ bound condition has the following property: $w_{it_i} \geq w_{it_i^*}$ and $e_{t_i i} \geq e_{t_i^* i}$ for all i always hold with at least one strict inequality, where t_i and t_i^* are treatments ex post assigned to i under the alternative design (p_{it}) and EXaM, respectively. Ordinal ex ante optimality is a stronger property that no other (p_{it}) satisfies the ϵ bound condition and that for all affine transformations f and g such that $\Sigma_t p_{it} f(w_{it}) \geq \Sigma_t p_{it}^*(\epsilon) f(w_{it})$ and $\Sigma_t p_{it} g(e_{ti}) \geq \Sigma_t p_{it}^*(\epsilon) g(e_{ti})$ for all i with at least one strict inequality.

erything else $(c_t^n, w_{it}^n, e_{ti}^n)$ may change as n increases. This modeling with a fixed number of treatments and an increasing number of subjects is consistent with real-world experiments with only a few treatments but with hundreds or thousands of subjects. Only for this section, for simplicity, I restrict w_{it}^n and e_{ti}^n to belong to finite sets W and E, respectively, in any problem along the sequence. Let $\epsilon^n \in [0, \bar{\epsilon}^n]$ (where $\bar{\epsilon}^n$ is $\bar{\epsilon}$ for the n-th problem) be the value of the bound parameter ϵ the experimenter picks for the n-th problem in the sequence.

To investigate the incentive structure in EXaM, imagine that subjects report their WTP to EXaM. EXaM then uses the reported WTP to compute treatment assignment probabilities. For the n-th problem in the sequence, let $p_i^{*n}(w_i, w_{-i}; \epsilon^n)$ be EXaM's treatment assignment probability vector for subject i when subjects report WTP (w_i, w_{-i}) where $w_{-i} \equiv (w_j)_{j \neq i}$. I extend this notation to the case where other subjects' WTP reports are random:

$$p_i^{*n}(w_i, F_w; \epsilon^n) \equiv \int_{w_{-i} \in W^{n-1}} p_i^{*n}(w_i, w_{-i}; \epsilon^n) \times \Pr(w_{-i} | w_{-i} \sim_{iid} F_w) dw_{-i}.$$

Here $\Pr(w_{-i} | w_{-i} \sim_{iid} F_w)$ denotes the probability that the reported WTP vector w_{-i} is realized from $n-1$ iid draws from the distribution $F_w \in \Delta W$ where ΔW is the set of full support distributions over the WTP space W. This concept allows me to define and state an asymptotic incentive compatibility property for EXaM.

Proposition 3 (Incentive)**.** *EXaM with WTP reporting is asymptotically incentive compatible, i.e., for any sequence of experimental design problems with any ϵ^n's in $[0, \bar{\epsilon}^n]$, any $F_w \in \Delta W$, any $\delta > 0$, there exists n_0 such that, for any $n \geq n_0$, any subject i, any true and manipulated WTP values $w_i, w_i' \in W$, I have*

$$\Sigma_t p_{it}^{*n}(w_i, F_w; \epsilon^n) \times w_{it} \geq \Sigma_t p_{it}^{*n}(w_i', F_w; \epsilon^n) \times w_{it} - \delta.$$

Proposition 3 says that for large enough experimental design problems, EXaM approximately incentivizes every subject to report her true WTP. It is an asymptotic theoretical result assuming the number of subjects growing to infinity. As additional support for incentive compatibility, Section 7.3 shows that EXaM is close to incentive compatible in my empirical application only with a modest finite number of subjects. This suggests asymptotic Proposition 3 is relevant even for real-scale problems.

5 Information: Econometrics of EXaM

Despite the welfare and incentive properties, the experimenter can extract as much information with EXaM as with RCT. To spell it out, take any experimental design problem as given. Suppose the experimenter is interested in the causal effect of each treatment on an outcome Y. Following the standard potential outcome framework for causal inference (Imbens and Rubin, 2015), let $Y_i(t)$ denote subject i's potential outcome that would be observed if subject i receives treatment t. Let D_{it} be the binary indicator that subject i is ex post assigned to treatment t. The observed outcome is written as $Y_i \equiv \Sigma_t D_{it} Y_i(t)$. While $Y_i(t)$ is assumed to be fixed, D_{it} and Y_i are random variables the distributions of which depend on the experimenter's choice of an experimental design. Let $Y \equiv (Y_i)$, $D_i \equiv (D_{it})_t$, and $D \equiv (D_i)$.

To compare EXaM and RCT in terms of their causal inference performance, I need to specify how each design draws a deterministic treatment assignment D_{it} from its assignment probabilities. For notational simplicity, assume that $p_t n_p$ is an integer for every t and p, where $n_p \equiv \Sigma_{i=1}^n 1\{p_i^*(\epsilon) = p\}$ and p_t is the element of p corresponding to treatment t. Appendix A.1.1 generalizes the definition and argument below to a general setting where $p_t n_p$ is any real number. Consider the following method of drawing a deterministic treatment assignment.

Definition 2 (EXaM Continued).

(3) Starting from the end of Definition 2 in Section 3.2, draw a treatment assignment from $p_{it}^*(\epsilon)$ as follows. For each propensity vector p,

- Step 1: Uniformly randomly pick $p_{t_0} n_p$ subjects from $\{i | p_i^*(\epsilon) = p\}$ and assign them to t_0.

For each subsequent step $k = 1, ..., m$,

- Step k: From the remaining $n_p - \Sigma_{t=t_0}^{t_{k-1}} p_t n_p$ subjects, uniformly randomly pick $p_{t_k} n_p$ subjects and assign them to t_k.

I also assume RCT to draw a deterministic treatment assignment by a special case of the above method, i.e., uniformly drawing any D_{it} satisfying $\Sigma_i D_{it} = c_t$ for each t and $\Sigma_t D_{it} = 1$ for each i.

With these preliminaries at hand, let θ be any parameter of interest of the distribution of potential outcomes $Y_i(t)$'s. Formally, θ is any mapping $\theta : \mathbb{R}^{n \times (m+1)} \to \mathbb{R}$ that maps each possible value of $(Y_i(t))_{it}$ into the corresponding value of the parameter. For example, θ

may be the average treatment effect of treatment t_1 over control t_0, $\frac{\Sigma_{i=1}^{n}(Y_i(t_1)-Y_i(t_0))}{n}$. I say parameter θ is *unbiasedly estimable with experimental design* $p \equiv (p_{it})_{it}$ *and a simple estimator* if there exists an estimator $\hat{\theta}(Y,D)$ such that

$$E(\hat{\theta}(Y,D)|p) = \theta$$

and $\hat{\theta}(Y,D)$ can be written as

$$\hat{\theta}(Y,D) = \Sigma_i f(Y_i, D_i) + \Sigma_t \Sigma_p \Sigma_{p'} g_{tpp'} \hat{\mu}_p(t) \hat{\mu}_{p'}(t)$$

where $\hat{\mu}_p(t) = \frac{\Sigma_{i:p_i=p} D_{it} Y_i}{p_t n_p}$. Here $E(\cdot|p)$ is expectation with respect to the distribution of D_{it} induced by experimental design p.[25] A measure of the informativeness of an experimental design is the set of parameters unbiasedly estimable with the experimental design and a simple estimator. In terms of this measure, EXaM turns out to be at least as informative as RCT.

Proposition 4 (Information)**.** *Parameter θ is unbiasedly estimable with EXaM $p^*_{it}(\epsilon)$ with any $\epsilon > 0$ and a simple estimator if θ is unbiasedly estimable with RCT p^{RCT}_{it} and a simple estimator.*[26]

Many parameters, such as the average treatment effect (ATE) and the mean and variance of potential outcomes, are known to be unbiasedly estimable with RCT and a simple estimator. Proposition 4 implies all of such parameters are also unbiasedly estimable with EXaM.

I use ATE to illustrate the intuition for and implementation of Proposition 4. EXaM makes all subjects share the same budget constraint and tie-breaking rule. As a result, if subjects share the same predicted effects and WTP, these subjects solve the same utility maximization problem and get the same vector of treatment assignment probabilities. EXaM therefore produces treatment assignment that is independent from (unconfounded by) potential outcomes conditional on predicted effects and WTP, which are observable to

[25] Since $Y_i(t)$ is constant and $Y_i = \Sigma_t D_{it} Y_i(t)$, the only source of randomness in $\hat{\theta}(Y,D)$ is randomness in D_{it}. Also, I allow estimator $\hat{\theta}(Y,D)$ to implicitly use known elements of the experimental design problem such as WTP w_{it}, predicted effects e_{ti}, and treatment assignment probabilities p_{it}. I do not allow $\hat{\theta}(Y,D)$ to use unknown elements, especially potential outcomes.

[26] This measure of informativeness induces an order over experimental designs. A design p is more informative than another design q if the set of parameters unbiasedly estimable with design p and a simple estimator includes the set of parameters unbiasedly estimable with design q and a simple estimator. This order is a relaxation of Blackwell's order (Blackwell and Girshick, 1954). That is, if p is more informative than q in Blackwell's sense, p is also more informative than q in my sense (but not vice versa). As a consequence, my order allows me to compare more experimental designs than Blackwell's does. For example, EXaM and RCT are comparable in my sense but incomparable in Blackwell's sense.

the experimenter:

$$(Y_i(t))_t \perp\!\!\!\perp D_i | (e_{ti}, w_{it})_t. \tag{1}$$

With this conditional independence, EXaM fits into causal inference with stratified experiments, selection-on-observables, and the propensity score, i.e., treatment assignment probabilities conditional on observables (see Imbens and Rubin (2015) for an overview). In particular, conditional independence (1) and Rosenbaum and Rubin (1983)'s result imply that the same conditional independence holds conditional on the propensity score, which EXaM computes as $p_i^*(\epsilon) \equiv (p_{it}^*(\epsilon))_t$ and again known to the econometrician:

$$(Y_i(t))_t \perp\!\!\!\perp D_i | p_i^*(\epsilon) \tag{2}$$

This conditionally independent treatment assignment allows the experimenter to unbiasedly estimate the conditional average treatment effects of each t over t_0 conditional on observable propensity scores $p_i^*(\epsilon)$,

$$\frac{\Sigma_{i=1}^n 1\{p_i^*(\epsilon) = p\}(Y_i(t) - Y_i(t_0))}{\Sigma_{i=1}^n 1\{p_i^*(\epsilon) = p\}} \text{ for each } p,$$

which I denote by $CATE_{pt}$. These conditional-on-the-propensity-score effects are a version of Marginal Treatment Effects (Björklund and Moffitt, 1987; Heckman and Vytlacil, 2005). Marginal Treatment Effects are therefore identifiable with EXaM's data.[27] By summing up such marginal or conditional effects, the experimenter can also back out the (unconditional) ATE, the single most important causal object identified and estimated by RCT. That is, with weights $\delta_p \equiv \frac{\Sigma_{i=1}^n 1\{p_i^*(\epsilon) = p\}}{n}$, I use $CATE_{pt}$'s to get ATE as follows:

$$\Sigma_p \delta_p CATE_{pt} = \frac{\Sigma_{i=1}^n (Y_i(t_1) - Y_i(t_0))}{n}.$$

Importantly, the key conditional independence properties (1) and (2) hold regardless of whether e_{ti} and w_{it} coincide with the true treatment effects and WTP. In this sense, like RCT, EXaM's informational virtue is robust to the experimenter's any misspecification

[27] To see this, as in Heckman and Vytlacil (2005), focus on an experimental design problem with only one treatment t_1 compared to the control t_0. Given EXaM's assignment probability $p_{it_1}^*(\epsilon)$, let $r_i \sim U[0,1]$ with $r_i \perp\!\!\!\perp (Y_i(t_0), Y_i(t_1))$, $Z_i = 1 - r_i$, and $V_i = 1 - p_{it_1}^*(\epsilon)$. Write the treatment assignment as

$$D_{it_1} = 1\{r_i \leq p_{it_1}^*(\epsilon)\} = 1\{1 - r_i \geq 1 - p_{it_1}^*(\epsilon)\} = 1\{Z_i \geq V_i\}.$$

Note that $E(1\{Z_i \geq V_i\}) = p_{it_1}^*(\epsilon)$ as desired. This model is a special case of Heckman-Vytlacil's model with local instrumental variable Z_i because Z_i is independent of $(Y_i(t_0), Y_i(t_1), V_i)$ by construction while V_i can be correlated with $(Y_i(t_0), Y_i(t_1))$. As a result, Heckman and Vytlacil (2005)'s method allows the experimenter to identify Marginal Treatment Effects with EXaM's data.

about predicted effects and WTP.

The above estimability argument motivates a regression strategy to estimate ATE with EXaM's data. As a warm-up, focus on $\{i|p_i^*(\epsilon) = p\}$, the subpopulation of subjects with propensity vector p, and consider this regression on the subpopulation:

$$Y_i = \alpha_p + \Sigma_{t=t_1}^{t_m} \beta_{pt} D_{it} + \epsilon_i, \tag{3}$$

By conditional independence property (2), for each treatment $t \neq t_0$, OLS estimate $\hat{\beta}_{pt}$ from this regression is unbiased for $CATE_{pt}$. I then aggregate the resulting estimates $\hat{\beta}_{pt}$'s into $\Sigma_p \delta_p \hat{\beta}_{pt}$, which I denote by $\hat{\beta}_t$. This estimator unbiasedly estimates the average treatment effect with a variance characterized below.

Proposition 5 (Bias and Variance)**.** *Suppose that the data-generating process is EXaM $p^*(\epsilon) \equiv (p_{it}^*(\epsilon))_{it}$ with any $\epsilon > 0$. $\hat{\beta}_t$ is an unbiased estimator of the average treatment effect. In particular,*

$$E(\hat{\beta}_t | p^*(\epsilon)) = \frac{\Sigma_{i=1}^n (Y_i(t_1) - Y_i(t_0))}{n} \tag{4}$$

$$Var(\hat{\beta}_t | p^*(\epsilon)) = \sum_p \delta_p^2 \Big(\frac{S_{pt}^2}{p_t n_p} + \frac{S_{pt_0}^2}{p_{t_0} n_p} - \frac{S_{ptt_0}^2}{n_p} \Big). \tag{5}$$

where $\bar{Y}_p(t) \equiv \dfrac{\sum_{i:p_i^*(\epsilon)=p} Y_i(t)}{n_p}$ is the mean of $Y_i(t)$ in the subpopulation with propensity p, $S_{pt}^2 \equiv \dfrac{\sum_{i:p_i^*(\epsilon)=p} (Y_i(t) - \bar{Y}_p(t))^2}{n_p - 1}$ is the variance of $Y_i(t)$ in the subpopulation, and $S_{ptt'}^2 \equiv \dfrac{\sum_{i:p_i^*(\epsilon)=p} (Y_i(t) - Y_i(t') - (\bar{Y}_p(t) - \bar{Y}_p(t')))^2}{n_p - 1}$ is the variance of $Y_i(t) - Y_i(t')$ in the same subpopulation.

Alternatively, empirical researchers may prefer a single regression controlling for propensity vectors:

$$Y_i = \Sigma_{t=t_1}^{t_m} b_t D_{it} + \Sigma_{t=t_1}^{t_m} c_t p_{it}^*(\epsilon) + e_i, \tag{6}$$

producing an alternative estimator $\hat{b}_t$. As shown in the appendix, $\hat{b}_t$ is an unbiased estimator of a differently weighted treatment effects:

$$E(\hat{b}_t | p^*(\epsilon)) = \frac{\Sigma_p \lambda_p CATE_{pt}}{\Sigma_p \lambda_p} \text{ with weights } \lambda_p \equiv \delta_p p_t (1 - p_t). \tag{7}$$

5.1 Comparison of EXaM and RCT

Proposition 5 implies EXaM's ATE estimation is not only unbiased but also as precise as RCT's. With RCT's data, the most standard estimator of ATE of treatment t over control t_0 is the difference in the average outcome between subjects assigned to treatment t and those assigned to control t_0:

$$\hat{\beta}_t^{RCT} \equiv \frac{\sum_i Y_i D_{it}}{\sum_i D_{it}} - \frac{\sum_i Y_i D_{it_0}}{\sum_i D_{it_0}}.$$

This $\hat{\beta}_t^{RCT}$ is a special case of $\hat{\beta}_t$ when $p_{it}^*(\epsilon) = p_{it}^{RCT}$ and so unbiased by Proposition 5 or Imbens and Rubin (2015)'s Theorem 6.2. Proposition 5 also implies the variance of $\hat{\beta}_t^{RCT}$.[28]

Corollary 1 (Imbens and Rubin (2015)'s Theorem 6.2)**.**

$$E(\hat{\beta}_t^{RCT}|p^{RCT}) = \frac{\sum_{i=1}^n (Y_i(t_1) - Y_i(t_0))}{n} \tag{8}$$

$$V(\hat{\beta}_t^{RCT}|p^{RCT}) = \frac{S_t^2}{c_t} + \frac{S_{t_0}^2}{c_{t_0}} - \frac{S_{tt_0}^2}{n}, \tag{9}$$

where $S_t^2 \equiv \dfrac{\sum_i (Y_i(t) - \bar{Y}(t))^2}{n-1}$ *and* $S_{tt'}^2 \equiv \dfrac{\sum_i (Y_i(t) - Y_i(t') - (\bar{Y}(t) - \bar{Y}(t')))^2}{n-1}$.

How do the two variances, $V(\hat{\beta}_t|p^*(\epsilon))$ and $V(\hat{\beta}_t^{RCT}|p^{RCT})$, compare to each other? It depends on the distribution of potential outcomes and treatment assignment probabilities. In particular, EXaM may produce more precise ATE estimates ($V(\hat{\beta}_t|p^*(\epsilon)) < V(\hat{\beta}_t^{RCT}|p^{RCT})$) if heterogeneous potential outcomes are well correlated with EXaM's treatment assignment probabilities. The following example illustrates this possibility.

Example 1. Suppose there is only one treatment t_1, $n = 50$, and $c_{t_0} = c_{t_1} = 25$. The subjects are divided into two groups A and B of the same size based on their potential outcomes $Y_i(t_1)$ and $Y_i(t_0)$. For anybody in group A, I have $Y_i(t_1) = 1$ and $Y_i(t_0) = 1$. For anybody in group B, I have $Y_i(t_1) = 5$ and $Y_i(t_0) = 3$. Assume the experimenter correctly predicts treatment effects: $e_{t_1 i} = 0$ for every i in group A while $e_{t_1 i} = 2$ for group B. Let $w_{it_1} > 0$ for all subjects. EXaM with $\epsilon = 0$ gives the following treatment assignment probabilities[29]: $p_{it_1}^*(\epsilon) = 0.2$ for every i in group A while $p_{it_1}^*(\epsilon) = 0.8$ for group B. Under

[28]This application of Proposition 5 assumes that under RCT, every treatment assignment satisfying the capacity constraint (c_t) occurs equally likely.

[29]EXaM outputs these treatment assignment probabilities if I set $\alpha = -\dfrac{15b}{8}$, $\beta_{t_1} = 5b$, and $\beta_{t_0} = 0$ given an arbitrary b.

RCT, $p_{it_1}^{RCT} = 0.5$ for all subjects. Applying Proposition 5 and Corollary 1 to this example, I have

$$V(\hat{\beta}_t | p^*(\epsilon)) = 0 < \frac{9}{49} = V(\hat{\beta}_t^{RCT} | p^{RCT}).$$

This section demonstrates how the experimenter should use EXaM to estimate key treatment effects along with the fact that EXaM can estimate ATE with potentially smaller standard errors than RCT. To execute, verify, and quantify these observations, I implement EXaM in a particular empirical context and analyze its data with estimators like $\hat{b}_t$ and $\hat{\beta}_t$. Before doing so, I relate EXaM to existing experimental designs.

6 Comparison with Existing Designs

Classical Experimental Design

The traditional experimental design literature (Cox and Cochran (1992), Athey and Imbens (2017) Section 7) is as old as the very concept of randomized experiments. This literature focuses on how to design experiments for maximizing information measured by the power of testing the null hypothesis of no treatment effect, the mean squared error in treatment effect estimation, and so on. This focus on information continues in the modern literature on sequential and adaptive experimental design (Hahn et al., 2011). My interest is more in ethics and welfare.

Preference- and Response-adaptive Designs

With its interest in subject well-being measured by WTP and predicted effects, EXaM is closer to younger and smaller strands of the literature on preference- and response-adaptive experimental designs. Preference-adaptive designs reflect subject preferences into treatment assignment probabilities. For example, Randomized *Consent* Trials (originally proposed by Zelen (1979) and further advocated by Angrist and Imbens (1991)) randomize subjects into two groups. In one group, subjects are allowed to choose the treatment or the control based on their preferences. All subjects in the other group are assigned to the control.[30] Selective Trials by Chassang et al. (2012, 2015), where the treatment assignment probability

[30]In Randomized Consent Trial, treatment assignment is chosen by subjects and not by chance. Yet data from Randomized Consent Trial identifies causal effects since the initial random grouping works as a random instrumental variable for the non-random treatment assignment. Randomized Consent Trial is mathematically equivalent to the instrumental variable setting with one-sided noncompliance (Bloom, 1984). Crucially, Zelen (1979) proposed to intentionally introduce such noncompliance (choice) to improve subject well-being. See also Angrist and Imbens (1991)'s section 3.1 for a more refined discussion.

is increasing in the WTP for the treatment, can also be thought of as a preference-adaptive design.[31]

In complementary response-adaptive designs (reviewed by Hu and Rosenberger (2006) and Food and Drug Administration (2010)), the experimenter incorporates predicted treatment effects into treatment assignment probabilities. For example, the Randomized Play-the-Winner Rule (Zelen, 1969; Wei and Durham, 1978) more likely assigns a treatment to patients predicted to have better treatment effects.[32]

EXaM attempts to integrate preference- and response-adaptive designs into a unified design. With help from economic theory and causal inference, EXaM is formally shown to strike a best balance between WTP and predicted effects (Proposition 2) without hurting incentive compatibility (Proposition 3) and experimental information (Propositions 4 and 5).

Multi-Armed Bandit Algorithms

EXaM shares much of its spirit with Multi-Armed Bandit (MAB) algorithms in computer science, machine learning, and statistics (Gittins et al., 2011): Both MAB and EXaM attempt to strike a balance between exploration (information) and exploitation (subject or experimenter welfare). MAB algorithms are popular in the web industry, especially for online ads and recommendations (White, 2012). Among many differences between MAB and EXaM, MAB mostly ignores incentive issues. In contrast, EXaM is formally and empirically shown to be nearly incentive compatible.[33]

[31] Chassang et al. (2012) proposed Selective Trials not for respecting subject welfare but for obtaining more information about the effect of subjects' treatment usage intensity on outcomes. Here I reinterpret Selective Trials from my well-being perspective. Also, other examples of preference-adaptive designs are recent RCTs in development economics that elicit subject preferences for treatment (Ashraf et al., 2006; Cohen and Dupas, 2010; Ashraf et al., 2010; Kremer et al., 2011; Devoto et al., 2012; Dupas, 2014). Many of their experimental designs are implicitly preference-adaptive.

[32] The treatment assignment literature in econometrics (Manski, 2008) and medicine (Chakraborty and Moodie, 2013) attempts a related but distinct task of using experimental data to optimally assign treatment to maximize welfare alone. The treatment assignment literature also largely ignores incentive issues.

[33] There are a small number of recent studies on incentive-compatible multi-armed bandit mechanisms (Babaioff et al., 2014). It's hard to apply any of them to the experimental design problem because of particular structures they impose (like the focus on deterministic mechanisms and the availability of real monetary transfers).

7 Empirical Application

7.1 Background

My empirical test bed for EXaM is an application to a spring protection experiment in Kenya. Waterborne diseases, especially diarrhea, remain the second leading cause of death among children. Almost 20% of child deaths under age five (about 1.5 million each year) is due to diarrhea.[34] Indeed, the only quantitative United Nations Millennium Development Goal is in terms of "the proportion of the population without sustainable access to safe drinking water and basic sanitation," such as protected springs.[35] Yet there is controversy about the health impacts of indirect improvements like spring protection that may fall short of piping better water into the home. Experts argued that improving source water quality may have only limited effects since, for example, water is likely recontaminated in transport and storage. These arguments were made in the absence of any randomized experiment.

This controversy motivated Kremer et al. (2011) to analyze randomized spring protection conducted by an NGO (International Children Support) in Kenya in the mid 2000s. This experiment randomly selected springs to receive protection from the universe of 200 local unprotected springs (after some eligibility screening). As experimental subjects, the NGO selected at baseline and followed afterward a representative sample of about 1500 households that regularly used some of the 200 springs. Kremer et al. (2011) find that spring protection substantially improves source water quality and is moderately effective at improving household water quality after some recontamination. Diarrhea among children in treatment households falls by about a quarter of the baseline level. I call this real experiment "Kremer et al. (2011)'s experiment" and distinguish it from EXaM and RCT as formal concepts in my model.

I consolidate Kremer et al. (2011)'s experimental data and my methodological framework to empirically evaluate EXaM. With the language and notation of my model, experimental subjects are households in Kremer et al. (2011)'s survey sample. The protection of the spring each household uses and no protection are a single treatment t_1 and a control t_0, respectively. Each household i's WTP for better water access t_1 is denoted by w_{it_1}, which I estimate below. I also estimate the predicted treatment effect $e_{t_1 i}$ of spring protection t_1 on household i's child diarrhea outcome. Using this embedding, I implement EXaM and compare it with RCT in terms of welfare, information, and incentive.

[34] See UNICEF and WHO's joint document "Diarrhoea: Why Children Are Still Dying and What Can be Done," at `http://apps.who.int/iris/bitstream/10665/44174/1/9789241598415_eng.pdf`, retrieved October 2017.

[35] See `http://www.un.org/millenniumgoals/`, retrieved in October 2017

7.2 Heterogeneous Treatment Effects and WTP

Treatment Effects

For executing EXaM, I need to know w_{it_1} and $e_{t_1 i}$ to be substituted into EXaM. I estimate heterogeneous treatment effects $e_{t_1 i}$ of access to better water in the same way as Kremer et al. (2011). To describe it, it is useful to provide additional details of Kremer et al. (2011)'s experiment. The experimenter NGO aspired to eventually protect all the 200 springs but planned for the protection intervention to be phased in over four years due to financial and administrative constraints. In each round, a subset of springs were randomly picked to be protected. Figure I in Kremer et al. (2011) details the timeline of the experiment.

This experimental scheme legitimizes the following OLS regression at the (child i, spring j, survey round t)-level:

$$Y_{ijt} = (\phi_1 + \phi_2 X_i) T_{jt} + \alpha_i + \alpha_t + u_{ij} + \epsilon_{ijt}, \tag{10}$$

where Y_{ijt} is the binary outcome indicating that child i in a household drawing water from spring j at baseline has diarrhea in survey round t. X_i contains covariates of child i's household (baseline latrine or sanitation density, diarrhea prevention knowledge score, mother's years of education, child gender). T_{jt} is the binary treatment indicating that spring j is treated in survey round t. α_i, α_t, and u_{ij} are fixed effects. The treatment effect is $\phi_1 + \phi_2 X_i$ and is heterogeneous across subjects with different covariates X_i.

Estimates from the OLS regression (10) are in Table 3. The average treatment effect is about 5% absolute reduction or about 25% relative reduction in the diarrhea outcome Y_{ijt}. Households with better diarrhea prevention knowledge scores or mother education tend to have better treatment effects. This heterogeneity may be because such households more sensitively change their drinking and washing water in response to the availability of a clean water source.

I then use the OLS estimates to predict the treatment effect for each household i with $\hat{e}_{t_1 i} \equiv \hat{\phi}_1 + \hat{\phi}_2 X_i$, where $\hat{\phi}_1$ and $\hat{\phi}_2$ are OLS estimates of ϕ_1 and ϕ_2, respectively. Kremer et al. (2011)'s experiment randomized T_{jt} and gives its coefficient $\hat{e}_{t_1 i}$ an interpretation as a causal effect.[36]

Estimated treatment effects $\hat{e}_{t_1 i}$ exhibit significant heterogeneity, as illustrated in Figure 1a. In Figure 1a, I simulate $\hat{e}_{t_1 i}$ with parametric bootstrap from $N(\hat{e}_{t_1 i}, SE(\hat{e}_{t_1 i}))$ and show

[36] X_i contains only household-level covariates except for child gender. When computing household-level $\hat{e}_{t_1 i}$, I code a household-level indicator for whether each household has a boy and substitute the indicator into X_i. This way, it is justifiable to interpret $\hat{e}_{t_1 i}$ as household-level predicted effects.

the histogram of the simulated values. The standard deviation is about 3.5% (shown in Appendix Table A.8).[37]

WTP

I estimate heterogeneous WTP w_{it_1} for better water as follows. In the experimental target area, each household draws water from a water source the household chooses among multiple sources in the neighborhood. This fact motivates a discrete choice model of households' water source choices, in which households trade off water quality against other source characteristics such as proximity. This model, combined with exogenous variation in water quality generated by Kremer et al. (2011)'s experiment, produces revealed preference estimates of household valuations of the spring protection treatment.

Specifically, I use the following mixed or random-coefficient logit model (Train (2003), chapter 6):

$$U_{ijt} = (\beta_i + \gamma_1 X_i)T_{jt} - c_i D_{ij} + \delta_j + \epsilon_{ijt}, \tag{11}$$

where U_{ijt} is household i's utility from source j in survey round t and D_{ij} is household i's roundtrip distance to spring j. β_i and c_i are random preference coefficients assumed to be distributed according to normal and triangular distributions, respectively. I impose the triangular assumption for c_i in order to make sure every household prefers proximity. δ_j are spring-type fixed effects in the spirit of Berry et al. (1995) and attempt to capture the average preference for potentially unobserved spring type characteristics other than treatment T_{jt} and distance D_{ij}. ϵ_{ijt} is logit utility shocks iid according to the type I extreme value distribution with usual variance normalization to $\pi^2/6$. I estimate the model with data on households' spring choices (in the final survey round) and a standard maximum simulated likelihood method (Train (2003), chapter 10), which I detail in Appendix A.2.3.

The mixed logit preference estimates are in Table 4. Households have significant distaste for distance and significant preferences for protected treatment springs (other characteristics being equal). Not surprisingly, households with better diarrhea prevention knowledge scores or mother education tend to have stronger revealed preferences for the spring protection treatment. This heterogeneity is expected if such households are more conscious of water quality.

I then exploit the mixed logit estimates to estimate household i's WTP for treatment t_1 as $\hat{w}'_{it_1} \equiv \hat{\beta}_i + \hat{\gamma}_1 X_i$, where $\hat{\beta}_i$ and $\hat{\gamma}_1$ are mixed logit estimates of β_i and γ_1, respectively. For the random coefficient $\hat{\beta}_i$, I bootstrap it from its estimated distribution. The identification of $\hat{w}'_{it_1}$ is helped by Kremer et al. (2011)'s experimental variation in protection treatment T_{jt}

[37] More details about each estimation or simulation procedure in this section are in Appendix A.2.3.

since otherwise T_{jt} is likely correlated with unobserved spring characteristics ϵ_{ijt}, making it impossible to identify the WTP for water quality alone.

Since $\hat{w}'_{it_1}$ is in an elusive utility unit, I convert it into a more easily-interpreted measure in terms of time cost of water collection. To do that, I first compute $\hat{w}'_{it_1}/\hat{c}_i$, where $\hat{c}_i$ is the mixed logit estimate of c_i (the distaste coefficient on distance). Again, I bootstrap the random coefficient $\hat{c}_i$ from its estimated distribution. I then multiply it by each household's self-reported time cost of traveling for a unit of distance. This procedure gives me a time cost measure of WTP for the treatment, $\hat{w}_{it_1}$. This $\hat{w}_{it_1}$ is measured by workdays utility-equivalent to $\hat{w}'_{it_1}$.

Estimated WTP $\hat{w}_{it_1}$ is in Figure 1b, showing the histogram of simulated values of $\hat{w}_{it_1}$. The median WTP is about 25 workday-equivalent with substantial heterogeneity (shown in Appendix Table A.8).

While both WTP $\hat{w}_{it_1}$ and treatment effects $\hat{e}_{t_1 i}$ show sizable heterogeneity, there turns out to be only limited correlation between the two. See the joint density plot in Figure 1c, where there is a positive correlation between WTP $\hat{w}_{it_1}$ and treatment effects $\hat{e}_{t_1 i}$, but the magnitude of the correlation is small (the OLS coefficient is about 0.005). This means that WTP $\hat{w}_{it_1}$ and treatment effects $\hat{e}_{t_1 i}$ contain different types of information about subject welfare, suggesting the importance of respecting both WTP and predicted effects separately. It is what EXaM attempts to do, as I explain next.

7.3 EXaM vs RCT

Now imagine somebody is planning a new experiment for further investigating the same spring protection treatment. What experimental design should she use? Specifically, which is better between RCT and EXaM? A full-fledged comparison of experimental designs requires a meta experiment that randomly assigns different designs to many experimental studies. To circumvent the difficulty with such a meta experiment, I resort to an alternative approach exploiting the above WTP and treatment effect estimates. My approach is to bootstrap WTP and predicted effects from the estimated distributions of $\hat{w}_{it_1}$ and $\hat{e}_{t_1 i}$, then use the bootstrapped data to simulate EXaM, and finally compare EXaM with RCT in terms of welfare, information, and incentive properties.

Throughout, I fix the set of subjects and treatments as in Kremer et al. (2011)'s experiment. That is, there are 1540 households as subjects to be assigned either to the single water source protection treatment t_1 or the control t_0. Set the treatment capacity c_{t_1} to be the number of households assigned to the treatment t_1 in Kremer et al.'s experiment (663). The control capacity c_{t_0} is the number of the remaining households, 877(=1540-663). I set

the bound parameter ϵ to be 0.1.

To run EXaM, the remaining pieces of necessary information are WTP w_{it_1} and predicted effects $e_{t_1 i}$. I simulate WTP and predicted effects with parametric bootstrap from the estimated distribution of $\hat{w}_{it_1}$ and $\hat{e}_{t_1 i}$, i.e., the estimated statistical models (10) and (11). Each bootstrap sample is of the same size ($n = 1540$) as the original sample in Kremer et al. (2011)'s experiment. I have to resort to this model-based, parametric bootstrap instead of a nonparametric bootstrap since WTP and predicted effects are model-based objects not directly observed in the data.

I quantify the difference between EXaM and RCT mixing the above bootstrap with counterfactual experimental design simulations. In particular, after simulating $(e_{t_1 i}, w_{it_1})$, I compute treatment assignment probabilities $p^*_{it}(\epsilon)$ by running EXaM on the bootstrapped data along with other fixed parameters like the treatment capacity. The algorithm I use for executing EXaM is described in Appendix A.2.4. Appendix Figure A.1 shows that the resulting treatment assignment probabilities $p^*_{it_1}(\epsilon)$ with EXaM are often different from RCT's constant probability $p^{RCT}_{it_1}$.

Welfare

I then calculate two welfare measures for each household i:

$$w^*_i \equiv \Sigma_t p^*_{it}(\epsilon) w_{it} \text{ and } e^*_i \equiv \Sigma_t p^*_{it}(\epsilon) e_{ti}.$$

w^*_i and e^*_i are two ex ante welfare measures in terms of WTP and predicted effects, respectively, in my theoretical welfare analysis (Proposition 2). The simulation process for RCT is analogous except that the treatment assignment probability is fixed at $p^{RCT}_{it} \equiv c_t/n = 877/1540 = .56$. Note that this RCT is a hypothetical experimental design in lines with my Definition 1 and different from Kremer et al. (2011)'s experiment involving additional real-world complications. More details about simulation procedures in this section are in Appendix A.2.4.

I find EXaM to improve on RCT in terms of the welfare measures w^*_i and e^*_i, as Figure 2 shows. The figure draws the distribution of w^*_i and e^*_i over households and 1000 bootstrap samples. Among other things, the median of average WTP w^*_i for assigned treatments increases by about 5.7 workday-equivalent utilities or 60% under EXaM than under RCT. Another interpretation of this WTP improvement is about 23% of the average WTP for the treatment (about 25 workdays, as shown in Appendix Table A.8). Similarly, EXaM improves the median of e^*_i by about .6% absolute reduction or 30% reduction relative to RCT's level. This predicted effect benefit amounts to about 13% of the average treatment effect of the

spring protection found by Kremer et al. (2011). This suggests EXaM's welfare optimality (Proposition 2) is quantitatively and empirically relevant.

Information

Data from EXaM also allows me to obtain more or less the same econometric conclusion about treatment effects as RCT. I do the following procedure many times:

(1) Simulate $(e_{t_1 i}, w_{it_1})$, run EXaM to get treatment assignment probabilities $p^*_{it}(\epsilon)$, and use $p^*_{it}(\epsilon)$ to draw a final deterministic treatment assignment, denoted by $D_i \equiv 1\{i \text{ is ex post assigned to } t_1\}$.

(2) Simulate counterfactual or predicted outcome Y_i under D_i simulating the OLS model I estimate in the last section:

$$Y_i \equiv (\hat{\phi}_1 + \hat{\phi}_2 X_i) D_i + \hat{\alpha}_i + (\text{average of } \hat{\alpha}_t \text{ across all } t) + (\text{average of } \hat{u}_{ij} \text{ across all } j),$$

where objects with a hat mean estimates of the corresponding parameters in regression (10). I take the average of $\hat{\alpha}_t$'s and $\hat{u}_{ij}$'s to adapt regression (10) at the (i, j, t)-level to my counterfactual simulation setting at the household-i-level. Note that the above expression is not a regression but the definition of Y_i.

(3) Use the above simulated Y_i and D_i to estimate treatment effects with $\hat{b}_{OLS}$ from this OLS regression:

$$Y_i = a + bD_i + cp^*_{it_1}(\epsilon) + e_i,$$

where I control for propensity score $p^*_{it_1}(\epsilon)$ to make treatment assignment D_i conditionally random, as suggested by the discussion in Section 5. This regression is a stripped-down version of the regression strategy (6) in Section 5.

The procedure for RCT is analogous except that the treatment assignment probability is fixed at $p^{RCT}_{it} \equiv c_t/n$.

Causal inference with EXaM turns out to be as unbiased and precise as that with RCT. Figure 3 plots the distribution of the resulting treatment effect estimates $\hat{b}_{OLS}$ and its p values (both robust and non-robust) over 1000 simulations for each experimental design. Panel 3a shows that consistent with Propositions 4 and 5, the median and mean of $\hat{b}_{OLS}$ for EXaM are indistinguishable from that under RCT. Both experimental designs successfully recover Kremer et al. (2011)'s average treatment effect estimate (4.5% reduction in diarrhea; recall column 1 in Table 3).

Perhaps more importantly, the randomization distribution of $\hat{b}_{OLS}$ for EXaM has a similar standard deviation as that for RCT. This means that the two experimental designs produce similar exact, finite-sample standard errors in their ATE estimates. Variations of this observation are Panels 3b and 3c, which show the randomization distribution of the p values for their ATE estimates (robust and non-robust, respectively). RCT is slightly more likely to produce smaller p values than EXaM, but their median p values are .015 for RCT and .021 for EXaM, meaning that both EXaM and RCT detect a significant average treatment effect for a majority of cases. Overall, EXaM appears to succeed in its informational mission of eliminating selection bias and recovering the average treatment effect precisely enough.

Incentive

Finally, EXaM's WTP benefits can be regarded as welfare-relevant only if EXaM provides subjects with incentives to reveal their true WTP. I conclude my empirical analysis with an investigation of the incentive compatibility of EXaM. I repeat the following procedure many times.

(1) As before, simulate $(e_{t_1 i}, w_{it_1})$ and run EXaM to get treatment assignment probabilities $p^*_{it}(\epsilon)$.

(2) Randomly pick one subject j as a WTP manipulator and one potential WTP manipulation w'_{jt_1} by j. I choose the manipulator j uniformly randomly among all subjects while w'_{jt_1} is from $N(w_{jt_1}, 100)$. Run EXaM on the simulated data in step (1) but with the WTP manipulation w'_{jt_1} to get treatment assignment probabilities $p'_{it}(\epsilon)$

(3) Compute the true WTP gain from the manipulation w'_{jt_1}: $\Delta w \equiv \Sigma_t p'_{it}(\epsilon) w_{jt} - \Sigma_t p^*_{it}(\epsilon) w_{jt}$.

As desired and consistent with Proposition 3, EXaM is found to give subjects little incentive for WTP misreporting. Figure 4 shows this by drawing the distribution of Δw over 1000 simulations. The WTP gain Δw from misreporting is mostly negative and well below zero on average. This suggests that EXaM may provide subjects with stronger average incentives for truthful WTP reporting than RCT does (because subjects in RCT are indifferent among all possible WTP reports). EXaM may therefore be better at eliciting reliable WTP data.

8 Discussion

8.1 Uncertainty in Predicted Effects and Preferences

The experimenter's information about preferences and predicted effects may be uncertain and probabilistic. What experimental design should the experimenter use with uncertain preferences and predicted effects? An *uncertain experimental design problem* consists of experimental subjects, treatments, treatment capacities, and the following objects.

- Each subject i's *preference or WTP* $\tilde{w}_{it}$ for treatment t where $\tilde{w}_{it}$ is a random variable.
- Each treatment t's *predicted treatment effect* $\tilde{e}_{ti}$ for subject i where $\tilde{e}_{ti}$ is a random variable.

$\tilde{w}_{it}$ and $\tilde{e}_{ti}$ are the experimenter's statistical perceptions about WTP and predicted treatment effect, respectively. I normalize $\tilde{e}_{ti}$ and $\tilde{w}_{it}$ by assuming $\tilde{e}_{t_0 i} = \tilde{w}_{it_0} = 0$ with probability 1 for every subject i. Denote $w_{it} \equiv E(\tilde{w}_{it})$ and $e_{ti} \equiv E(\tilde{e}_{ti})$ where each expectation is with respect to the distribution of $\tilde{w}_{it}$ and $\tilde{e}_{ti}$, respectively.

When I apply EXaM to (w_{it}, e_{ti}), the resulting EXaM nests RCT, is efficient with respect to (w_{it}, e_{ti}), is approximately incentive compatible, and is as informative as RCT in the same senses as in Propositions 1-4, respectively.

8.2 Ordinal Predicted Effects and Preferences

The experimenter's information about preferences and predicted effects may be ordinal. What experimental design should the experimenter use with ordinal preferences and predicted effects? An *ordinal experimental design problem* consists of experimental subjects, treatments, treatment capacities, and the following objects.

- Each subject i's *ordinal preference* $\succsim_i$ for treatment t where $t \succsim_i t'$ means subject i weakly prefers treatment t over t'. $\succsim_i$ may involve ties and indifferences.
- Each treatment t's *ordinal predicted treatment effect* $\succsim_t$ for subject i where $i \succsim_t i'$ means treatment t is predicted weakly more effective for subject i than for subject i'. Again, $\succsim_t$ may involve ties and indifferences.

I consider the following adaptation of EXaM to this ordinal experimental design problem.

Definition 3 (Ordinal EXaM)**.** (1) Create any cardinal WTP w'_{it} of each subject i for each treatment t so that $w'_{it} > w'_{it'}$ if and only if $t \succ_i t'$.

(2) Create any cardinal predicted effect of each treatment t for each subject i so that $e'_{ti} > e'_{t'i}$ if and only if $i \succ_t i'$.

(3) Run EXaM (as defined in Definition 2) on (w'_{it}, e'_{ti}) to get treatment assignment probabilities $p^{*o}_{it}(\epsilon)$

Ordinal EXaM nests RCT, is approximately incentive compatible, and is as informative as RCT in the same senses as in Propositions 1, 3, and 4, respectively. Moreover, ordinal EXaM has the following nice welfare property with respect to ordinal preferences and predicted effects.

Proposition 6. *$p^{*o}_{it}(\epsilon)$ is ordinally efficient in the following sense. There is no other experimental design (p_{it}) with $p_{it} \in [\epsilon, 1-\epsilon]$ for all subject i and treatment t and such that for all cardinal WTP w_{it} consistent with ordinal $\succsim_i$ and all cardinal predicted effects e_{ti} consistent with ordinal $\succsim_t$, I have*

$$\Sigma_t p_{it} w_{it} \geq \Sigma_t p^{*o}_{it}(\epsilon) w_{it} \text{ and } \Sigma_t p_{it} e_{ti} \geq \Sigma_t p^{*o}_{it}(\epsilon) e_{ti}$$

for all i with at least one strict inequality.

9 Takeaway and Future Directions

Motivated by the high-stakes nature of many RCTs, I propose an experimental design dubbed as Experiment-as-Market (EXaM). EXaM is a solution to a hybrid experimental-design-as-market-design problem of maximizing subjects' welfare subject to the constraint that the experimenter must produce as much information and incentives as RCT (Propositions 2-4). These properties are then verified and quantified in an application to a water source protection experiment. Taken together, the body of evidence suggests empirical support for the idea that EXaM improves subject well-being with little or no information and incentive costs. The demonstrated benefits are conservative in that it does not incorporate potential additional benefits from EXaM for improving recruitment, compliance with assigned treatment, and attrition (recall the discussion in Section 2).

This paper takes a step toward introducing welfare and ethics into the randomized experiment landscape. This opens the door to several open questions. For example, practically, the most crucial step is to implement EXaM in the field. In order to make EXaM and other designs workable in practice, it is also important to empirically understand the size and inner working of the clinical trial industry. A brief analysis in Section 2 is such an effort.

Econometrically and theoretically, this paper's analysis is simplistic in many respects, and a variety of extensions are in order. Key extensions include analyzing EXaM in an instrumental variable setting where subjects may not comply with treatment assignment; analyzing experimental designs with endogenous subject participation and dropout; introducing monetary compensation and other contracts like informed consent; analyzing EXaM's dynamic or sequential properties; optimally choosing sample size and treatment definitions (in addition to designing treatment assignment probabilities given the sample size and treatment definition); considering information frictions and psychological elements in patient preferences; and analyzing games among experimenters with experimental design as an action or strategy. It is also intriguing to use the EXaM framework to analyze external validity of causal inference with different experimental designs. I leave these challenging directions for future research.

References

Abdulkadiroğlu, Atila, Joshua Angrist, Yusuke Narita, and Parag A. Pathak, "Research Design Meets Market Design: Using Centralized Assignment for Impact Evaluation," *Econometrica*, 2017, *85* (5), 1373–1432.

Angelucci, Manuela and Daniel Bennett, "The Marriage Market for Lemons: HIV Testing and Marriage in Rural Malawi," 2017. Working Paper.

Angrist, Joshua and Guido Imbens, "Sources of Identifying Information in Evaluation Models," 1991. Working Paper.

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

_ , James Berry, and Jesse M Shapiro, "Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia," *American Economic Review*, 2010, *100* (5), 2383–2413.

Athey, Susan and Guido W Imbens, "The Econometrics of Randomized Experiments," *Handbook of Economic Field Experiments*, 2017, *1*, 73–140.

Azevedo, Eduardo and Eric Budish, "Strategyproofness in the Large," 2017. Working Paper.

Babaioff, Moshe, Yogeshwer Sharma, and Aleksandrs Slivkins, "Characterizing Truthful Multi-armed Bandit Mechanisms," *SIAM Journal on Computing*, 2014, *43* (1), 194–230.

Baicker, Katherine, Sarah L Taubman, Heidi L Allen, Mira Bernstein, Jonathan H Gruber, Joseph P Newhouse, Eric C Schneider, Bill J Wright, Alan M Zaslavsky, and Amy N Finkelstein, "The Oregon Experiment — Effects of Medicaid on Clinical Outcomes," *New England Journal of Medicine*, 2013, *368* (18), 1713–1722.

Banerjee, Abhijit and Esther Duflo, *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*, PublicAffairs, 2012.

Berry, Steven, James Levinsohn, and Ariel Pakes, "Automobile Prices in Market Equilibrium," *Econometrica*, 1995, pp. 841–890.

Björklund, Anders and Robert Moffitt, "The Estimation of Wage Gains and Welfare Gains in Self-selection Models," *Review of Economics and Statistics*, 1987, pp. 42–49.

Blackwell, David and MA Girshick, *Theory of Games and Statistical Decisions*, Dover Publications, 1954.

Bloom, Howard S, "Accounting for No-shows in Experimental Evaluation Designs," *Evaluation Review*, 1984, *8* (2), 225–246.

Bó, Ernesto Dal, Frederico Finan, and Martín A Rossi, "Strengthening State Capabilities: The Role of Financial Incentives in the Call to Public Service," *Quarterly Journal of Economics*, 2013, *128* (3), 1169–1218.

Bogomolnaia, Anna and Hervé Moulin, "A New Solution to the Random Assignment Problem," *Journal of Economic theory*, 2001, *100* (2), 295–328.

Budish, Eric, Gérard P Cachon, Judd B Kessler, and Abraham Othman, "Course Match: A Large-Scale Implementation of Approximate Competitive Equilibrium from Equal Incomes for Combinatorial Allocation," *Operations Research*, 2016.

_ , Yeon-Koo Che, Fuhito Kojima, and Paul Milgrom, "Designing Random Allocation Mechanisms: Theory and Applications," *American Economic Review*, 2013, *103* (2), 585–623.

Chakraborty, Bibhas and Erica EM Moodie, *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*, Springer, 2013.

Chan, Tat Y and Barton H Hamilton, "Learning, Private Information, and the Economic Evaluation of Randomized Experiments," *Journal of Political Economy*, 2006, *114* (6), 997–1040.

Chassang, Sylvain, Erik Snowberg, Ben Seymour, and Cayley Bowles, "Accounting for Behavior in Treatment Effects: New Applications for Blind Trials," *PloS One*, 2015, *10* (6), e0127227.

_ , Gerard Padró i Miquel, and Erik Snowberg, "Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments," *American Economic Review*, 2012, *102* (4), 1279–1309.

Chilvers, Clair, Michael Dewey, Katherine Fielding, Virginia Gretton, Paul Miller, Ben Palmer, David Weller, Richard Churchill, Idris Williams, Navjot Bedi et al., "Antidepressant Drugs and Generic Counselling for Treatment of Major Depression in Primary Care: Randomised Trial with Patient Preference Arms," *British Medical Journal*, 2001, *322* (7289), 772.

Cohen, Jessica and Pascaline Dupas, "Free Distribution or Cost-sharing? Evidence from a Randomized Malaria Prevention Experiment," *Quarterly Journal of Economics*, 2010, pp. 1–45.

Cohen, Myron S, Ying Q Chen, Marybeth McCauley, Theresa Gamble, Mina C Hosseinipour, Nagalingeswaran Kumarasamy, James G Hakim, Johnstone Kumwenda, Beatriz Grinsztejn, Jose HS Pilotto et al., "Prevention of HIV-1 Infection with Early Antiretroviral Therapy," *New England Journal of Medicine*, 2011, *365* (6), 493–505.

Cox, Gertrude M and WG Cochran, *Experimental Designs*, Wiley, 1992.

DellaVigna, Stefano and Devin Pope, "Predicting Experimental Results: Who Knows What?," 2016. Working Paper.

Devoto, Florencia, Esther Duflo, Pascaline Dupas, William Parienté, and Vincent Pons, "Happiness on Tap: Piped Water Adoption in Urban Morocco," *American Economic Journal: Economic Policy*, 2012, *4* (4), 68–99.

Duflo, Esther, Rachel Glennerster, and Michael Kremer, "Using Randomization in Development Economics Research: A Toolkit," *Handbook of Development Economics*, 2007, *4*, 3895–3962.

Dupas, Pascaline, "Short-run Subsidies and Long-run Adoption of New Health Products: Evidence from a Field Experiment," *Econometrica*, 2014, *82* (1), 197–228.

Epstein, Ronald M and Ellen Peters, "Beyond Information: Exploring Patients' Preferences," *Journal of American Medical Association*, 2009, *302* (2), 195–197.

Food and Drug Administration, "Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics," 2010. http://www. fda. gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm201790. pdf.

__ , "Patient Preference Information," 2016. https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM446680.pdf.

Freeman, Bradley D, Robert L Danner, Steven M Banks, and Charles Natanson, "Safeguarding Patients in Clinical Trials with High Mortality Rates," *American Journal of Respiratory and Critical Care Medicine*, 2001, *164* (2), 190–192.

Friedman, Lawrence M, Curt Furberg, David L DeMets, David M Reboussin, and Christopher B Granger, *Fundamentals of Clinical Trials*, Vol. 3, Springer, 1998.

Friedman, Milton, *Capitalism and Freedom*, University of Chicago Press, 1962.

Gaw, Allan, *Trial by Fire: Lessons from the History of Clinical Trials*, SA Press, 2009.

Gerber, Alan S and Donald P Green, *Field Experiments: Design, Analysis, and Interpretation*, WW Norton, 2012.

Gittins, John, Kevin Glazebrook, and Richard Weber, *Multi-armed Bandit Allocation Indices*, John Wiley & Sons, 2011.

Goldacre, Ben, *Bad Pharma: How Drug Companies Mislead Doctors and Harm Patients*, Macmillan, 2014.

Grant, Robert M, Javier R Lama, Peter L Anderson, Vanessa McMahan, Albert Y Liu, Lorena Vargas, Pedro Goicochea, Martín Casapía, Juan Vicente Guanira-Carranza, Maria E Ramirez-Cardich et al., "Preexposure Chemoprophylaxis for HIV Prevention in Men who Have Sex with Men," *New England Journal of Medicine*, 2010, *2010* (363), 2587–2599.

Gueron, Judith M and Howard Rolston, *Fighting for Reliable Evidence*, Russell Sage Foundation, 2013.

Hahn, Jinyong, Keisuke Hirano, and Dean Karlan, "Adaptive Experimental Design Using the Propensity Score," *Journal of Business and Economic Statistics*, 2011, *29* (1), 96–108.

Haushofer, Johannes and Jeremy Shapiro, "The Short-term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya," *Quarterly Journal of Economics*, 2016, *131* (4), 1973–2042.

He, Yinghua, Antonio Miralles, Marek Pycia, and Jianye Yan, "A Pseudo-Market Approach to Allocation with Priorities," *American Economic Journal: Microeconomics*, 2017.

Heckman, James J and Edward Vytlacil, "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 2005, *73* (3), 669–738.

Henshaw, RC, SA Naji, IT Russell, and AA Templeton, "Comparison of Medical Abortion with Surgical Vacuum Aspiration: Women's Preferences and Acceptability of Treatment.," *British Medical Journal*, 1993, *307* (6906), 714–717.

Hu, Feifang and William F Rosenberger, *The Theory of Response-adaptive Randomization in Clinical Trials*, Vol. 525, John Wiley & Sons, 2006.

Hylland, Aanund and Richard J. Zeckhauser, "The Efficient Allocation of Individuals to Positions," *Journal of Political Economy*, 1979, *87(2)*, 293–314.

Imbens, Guido W and Donald B Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge University Press, 2015.

Jackson, Matthew O, "Incentive Compatibility and Competitive Allocations," *Economics Letters*, 1992, *40* (3), 299–302.

Kass, Michael A, Dale K Heuer, Eve J Higginbotham, Chris A Johnson, John L Keltner, J Philip Miller, Richard K Parrish, M Roy Wilson, and Mae O Gordon, "The Ocular Hypertension Treatment Study: A Randomized Trial Determines That Topical Ocular Hypotensive Medication Delays or Prevents the Onset of Primary Open-angle Glaucoma," *Archives of Ophthalmology*, 2002, *120* (6), 701–713.

King, Michael, Irwin Nazareth, Fiona Lampe, Peter Bower, Martin Chandler, Maria Morou, Bonnie Sibbald, and Rosalind Lai, "Impact of Participant and Physician Intervention Preferences on Randomized Trials: A Systematic Review," *Journal of American Medical Association*, 2005, *293* (9), 1089–1099.

Kremer, Michael, Jessica Leino, Edward Miguel, and Alix Peterson Zwane, "Spring Cleaning: Rural Water Impacts, Valuation, and Property Rights Institutions," *Quarterly Journal of Economics*, 2011, *126* (1), 145–205.

Manski, Charles, *Identification for Prediction and Decision*, Cambridge: Harvard University Press, 2008.

Morgan, Steve, Paul Grootendorst, Joel Lexchin, Colleen Cunningham, and Devon Greyson, "The Cost of Drug Development: A Systematic Review," *Health Policy*, 2011, *100* (1), 4–17.

Narita, Yusuke, "(Non)Randomization: A Theory of Quasi-Experimental Evaluation of School Quality," 2016. Working Paper.

Preference Collaborative Review Group, "Patients' Preferences within Randomised Trials: Systematic Review and Patient Level Meta-analysis," *British Medical Journal*, 2008, *337*.

Roberts, Donald John and Andrew Postlewaite, "The Incentives for Price-taking Behavior in Large Exchange Economies," *Econometrica*, 1976, pp. 115–127.

Rosenbaum, Paul R. and Donald B. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrica*, 1983, pp. 41–55.

Scandinavian Simvastatin Survival Study Group and Others, "Randomised Trial of Cholesterol Lowering in 4444 Patients with Coronary Heart Disease: the Scandinavian Simvastatin Survival Study (4S)," *Lancet*, 1994, *344* (8934), 1383–1389.

Shamoo, Adil E and David B Resnik, *Responsible Conduct of Research*, Oxford University Press, 2009.

Sherman, Lawrence W and David Weisburd, "General Deterrent Effects of Police Patrol in Crime "Hot Spots": A Randomized, Controlled Trial," *Justice Quarterly*, 1995, *12* (4), 625–648.

Siroker, Dan and Pete Koomen, *A/B Testing: The Most Powerful Way to Turn Clicks into Customers*, John Wiley & Sons, 2013.

Stupp, Roger, Monika E Hegi, Warren P Mason, Martin J van den Bent, Martin JB Taphoorn, Robert C Janzer, Samuel K Ludwin, Anouk Allgeier, Barbara Fisher, Karl Belanger et al., "Effects of Radiotherapy with Concomitant and Adjuvant Temozolomide versus Radiotherapy Alone on Survival in Glioblastoma in a Randomised Phase III Study: 5-year Analysis of the EORTC-NCIC Trial," *Lancet Oncology*, 2009, *10* (5), 459–466.

Swift, Joshua K and Jennifer L Callahan, "The Impact of Client Treatment Preferences on Outcome: A Meta-analysis," *Journal of Clinical Psychology*, 2009, *65* (4), 368–381.

Train, Kenneth, *Discrete Choice Methods with Simulation*, Cambridge University Press, 2003.

UK Prospective Diabetes Study Group, "Tight Blood Pressure Control and Risk of Macrovascular and Microvascular Complications in Type 2 Diabetes: UKPDS 38," *British Medical Journal*, 1998, pp. 703–713.

Varian, Hal R, "Equity, Envy, and Efficiency," *Journal of Economic Theory*, 1974, *9* (1), 63–91.

Wei, LJ and S Durham, "The Randomized Play-the-winner Rule in Medical Trials," *Journal of the American Statistical Association*, 1978, *73* (364), 840–843.

White, John, *Bandit Algorithms for Website Optimization*, O'Reilly, 2012.

Writing Group for the Women's Health Initiative Investigators and Others, "Risks and Benefits of Estrogen plus Progestin in Healthy Postmenopausal Women: Principal Results from the Women's Health Initiative Randomized Controlled Trial," *Journal of American Medical Association*, 2002, *288* (3), 321–333.

Zarin, Deborah A, Tony Tse, Rebecca J Williams, and Sarah Carr, "Trial Reporting in ClinicalTrials.gov: The Final Rule," *New England Journal of Medicine*, 2016, *375* (20), 1998–2004.

Zelen, Marvin, "Play the Winner Rule and the Controlled Clinical Trial," *Journal of the American Statistical Association*, 1969, *64* (325), 131–146.

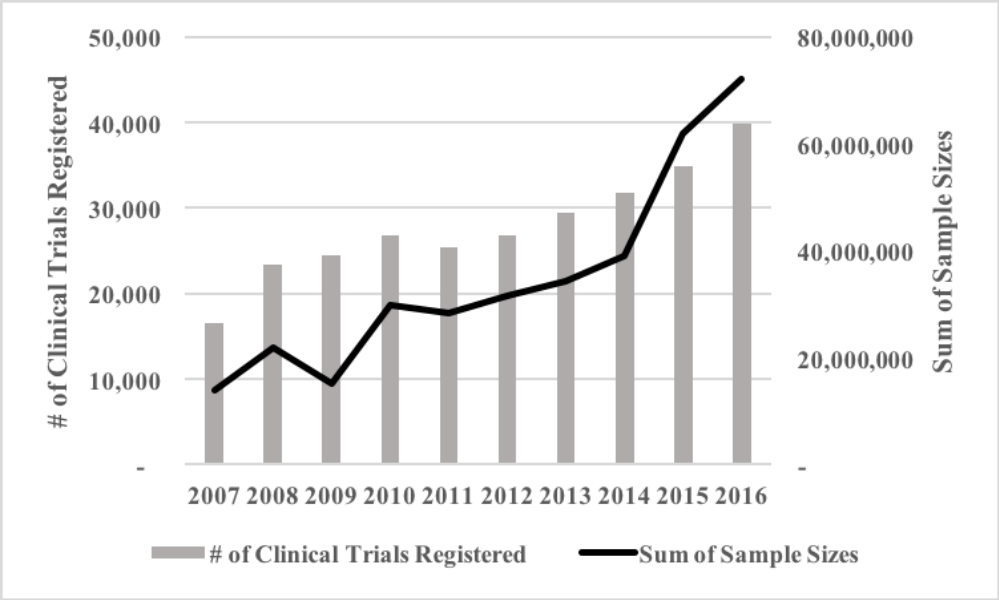
__ , "A New Design for Randomized Clinical Trials," *New England Journal of Medicine*, 1979, *300* (22), 1242–1245.

Table 1: Magnitude of a Part of the Clinical Trial Industry

(a) Registered Clinical Trials & Sample Sizes

	Sample Period 2007-2017 May
Total Number of Clinical Trials Registered	296,597
Sum of Sample Sizes	367,902,580

(b) Time Evolution



Notes: This table provides summary statistics of clinical trials registered in the WHO International Clinical Trials Registry Platform (ICTRP, `http://www.who.int/ictrp/en/`, retrieved in October 2017). The sample consists of clinical trials registered there between January 1st 2007 to May 30th 2017. I exclude trials with registered sample size larger than five millions. See Section 2 for discussions about this exhibit and Appendix A.2.1 for the detailed computational procedure. Additional results are in Appendix Tables A.1-A.3.

Table 2: A Selection of High-stakes Randomized Controlled Trials

(a) Medical Clinical Trials

	Treatment	Outcome	Treatment Effect
i	Cholesterol Lowering Drug	Mortality (in 5 Years)	30% Reduction
ii	Blood Pressure Control	Diabetes-Related Mortality (in 4.5 Years)	24% Reduction
iii	Medication to Reduce Interocular Pressre	Visual Field Abnormality (in 6 Years)	59% Reduction
iv	HIV Prevention Drug	HIV Infection Rate (in 1 Year)	44% Reduction
v	Antiretroviral Therapy	HIV Transmission Rate (in 1.7 Years)	96% Reduction
vi	Hormone Therapy	Coronary Heart Disease (in 5 Years)	29% Increase

(b) Social and Economic Experiments

	Treatment	Outcome	Treatment Effect
I	Unconditional Cash Transfers	Consumption (9 Months Later)	22% Increase
II	Police Patrol	Crime Calls	10% Reduction
III	HIV Testing	Fertility (in 2-3 Years)	18% Increase
IV	Health Insurance (Medicaid)	Depression Rate (in 2 Years)	30% Reduction
V	High Wage Job Offer	Offer Acceptance Rate	29% Increase

Notes: This table lists examples illustrating the high-stakes nature of certain RCTs. Following the convention in medical literature, treatment effects are measured relative to the average outcome in the control group, which I normalize to 100%. Every treatment effect is statistically significant at the 5% or lower level. The control is a placebo or the absence of any treatment unless otherwise noted. See the following references for the details of each RCT:

Panel (a) Study i: Scandinavian Simvastatin Survival Study Group and Others (1994)
Panel (a) Study ii: UK Prospective Diabetes Study Group (1998), where the control is less tight blood pressure control.
Panel (a) Study iii: Kass et al. (2002)
Panel (a) Study iv: Grant et al. (2010)
Panel (a) Study v: Cohen et al. (2011)
Panel (a) Study vi: Writing Group for the Women's Health Initiative Investigators and Others (2002)
Panel (b) Study I: Haushofer and Shapiro (2016)
Panel (b) Study II: Sherman and Weisburd (1995)
Panel (b) Study III: Angelucci and Bennett (2017)
Panel (b) Study IV: Baicker et al. (2013)
Panel (b) Study V: Dal Bó et al. (2013), where the control is a lower wage job offer.

See Section 2 for discussions about this table.

Table 3: OLS Regression Estimates of Heterogeneous Treatment Effects

Dependent Variable: Incidence of child diarrhea in past week						
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES						Main
Treatment	−0.045***	−0.045***	−0.046***	−0.044***	−0.054***	−0.052***
	(0.012)	(0.012)	(0.012)	(0.012)	(0.015)	(0.016)
Treatment * latrine density		−0.061				−0.036
		(0.069)				(0.066)
Treatment * diarrhea prevention			−0.012***			−0.010**
			(0.004)			(0.005)
Treatment * mother's education				−0.007**		−0.006**
				(0.003)		(0.003)
Treatment * boy					0.016	0.014
					(0.018)	(0.018)
Observations	6,750	6,750	6,750	6,742	6,670	6,662
Mean of dependent variable in comparison group				0.17		

Notes: This table shows OLS regression estimates of heterogeneous treatment effects of spring protection. Data from all four survey rounds (2004, 2005, 2006, 2007), sample restricted to children under age three at baseline (in 2004) and children born since 2004 in sample households. Diarrhea defined as three or more "looser than normal" stools within 24 hours at any time in the past week. Different columns differ in the set of baseline household characteristics interacted with the treatment indicator. The gender-age controls include linear and quadratic current age (by month), and these terms interacted with a gender indicator. I use specifications without additional controls. Stars *, **, and *** mean significance at 90%, 95%, and 99%, respectively, based on Huber-White robust standard errors clustered at the spring level. See Section 7.2 for the model description and discussions about this table.

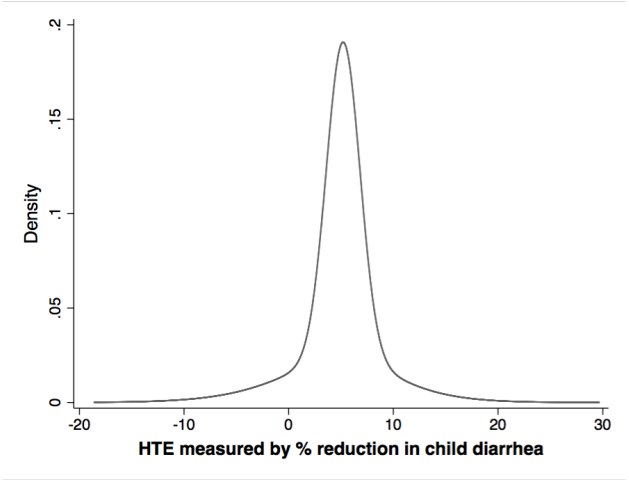
Table 4: Maximum Simulated Likelihood Estimates of Mixed Logit Model of Spring Choice

VARIABLES	(1)	(2)	(3)	(4)	(5) Main
Spring protection treatment indicator (Normal)					
Mean	2.205***	3.163***	2.999***	3.199***	3.370***
	(0.213)	(0.235)	(0.289)	(0.353)	(0.369)
S. D.	5.426***	5.702***	5.557***	6.282***	5.743***
	(0.298)	(0.291)	(0.405)	(0.394)	(0.313)
Treatment * latrine density	7.533***				2.585*
	(0.939)				(1.435)
Treatment * diarrhea prevention		1.080***			0.569***
		(0.104)			(0.104)
Treatment * mother's education			0.650***		0.623***
			(0.066)		(0.080)
Treatment * having a boy				−0.688	0.418
				(0.748)	(0.759)
Distance to source, minutes walk (Restricted triangular)					
Mean	−0.222***	−0.220***	−0.220***	−0.210***	−0.220***
	(0.010)	(0.010)	(0.010)	(0.009)	(0.010)
S. D.	0.222***	0.220***	0.220***	0.210***	0.220***
	(0.010)	(0.010)	(0.010)	(0.009)	(0.010)
Source type: borehole/piped	−1.079***	−1.047***	−1.055***	−1.070***	−1.051***
	(0.135)	(0.136)	(0.139)	(0.136)	(0.134)
Source type: well	−1.924***	−1.954***	−1.943***	−1.908***	−1.944***
	(0.137)	(0.131)	(0.134)	(0.132)	(0.132)
Source type: stream/river	−1.422***	−1.387***	−1.443***	−1.389***	−1.385***
	(0.144)	(0.141)	(0.148)	(0.146)	(0.141)
Source type: lake/pond	−0.312	−0.313	−0.333	−0.392	−0.295
	(0.269)	(0.273)	(0.274)	(0.371)	(0.312)
Number of observations (water collection choice situations)	53,427	53,427	53,427	53,427	53,427

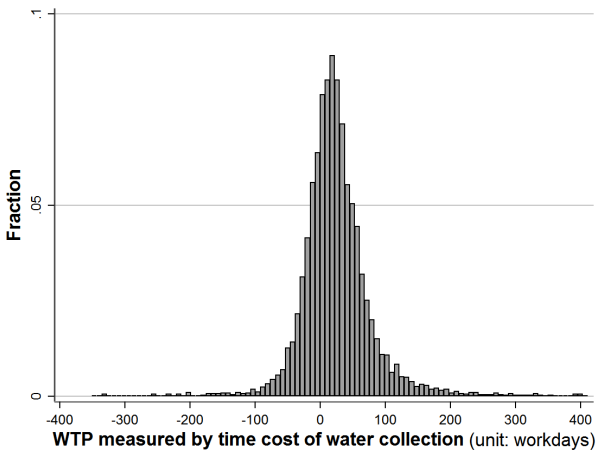
Notes: This table shows mixed logit estimates used for estimating heterogeneous WTP for the treatment. Each observation is a unique household-water source pair in one water collection trip recorded in the final round of household surveys (2007). The dependent variable is a multinomial indicator equaling 1 if the household chose the water source represented in the household-source pair. The omitted water source category is non-program springs outside the target area of the experiment. Different columns differ in the set of baseline household characteristics interacted with the treatment indicator. The indicator for the spring that each household used at baseline is in the models, but its coefficient estimate is not shown in the table. Stars *, **, and *** mean significance at 90%, 95%, and 99%, respectively. See Section 7.2 for the model description and discussions about this table. See Appendix A.2.3 for the estimation procedure to produce these estimates.

Figure 1: Treatment Effects and WTP for the Treatment

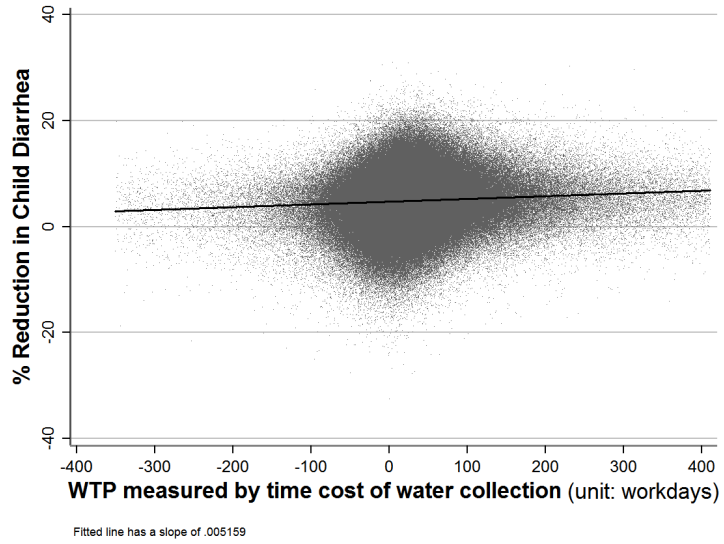
(a) Heterogeneity in Treatment Effects $\hat{e}_{t_1 i}$



(b) Heterogeneity in WTP $\hat{w}_{it_1}$



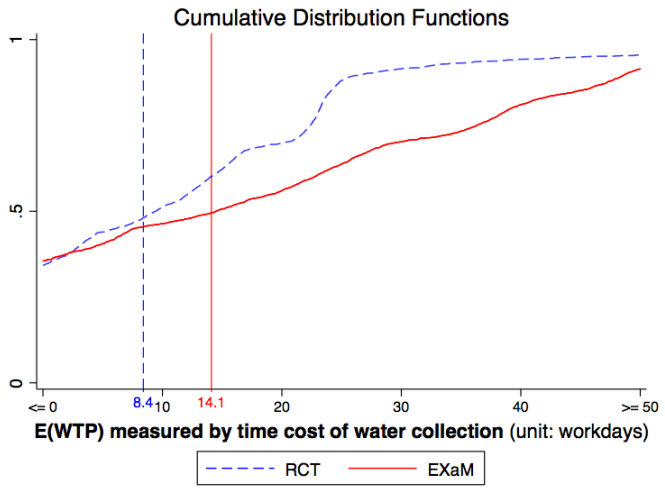
(c) Limited Correlation between Treatment Effects & WTP



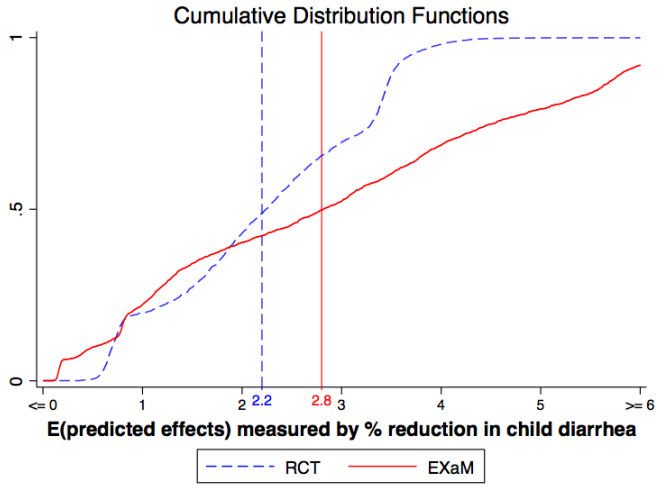
Notes: This figure shows the pattern of heterogeneity in estimated WTP $\hat{w}_{it_1}$ and predicted treatment effects $\hat{e}_{t_1 i}$. Panel 1a is about the predicted treatment effects $\hat{e}_{t_1 i}$ measured in percentage reduction in the incidence of child diarrhea in the past week, while Panel 1b is about WTP for the spring protection treatment $\hat{w}_{it_1}$, measured by time cost of water collection in the unit of workdays. I simulate values of $\hat{w}_{it_1}$ and $\hat{e}_{t_1 i}$ with parametric bootstrap from the main statistical specifications including all of the interactions between the treatment indicator and household characteristics (baseline latrine density, diarrhea prevention knowledge score, mother's years of education, and having a boy in the household). Panel 1c demonstrates the correlation between WTP $\hat{w}_{it_1}$ and predicted treatment effects $\hat{e}_{t_1 i}$. In order to emphasize visibility, I focus on the three standard deviations around the mean. See Section 7.2 for discussions about this figure. See Appendix A.2.3 for the detailed computational procedure.

Figure 2: EXaM vs RCT: Welfare

(a) Average WTP for Assigned Treatments w_i^*



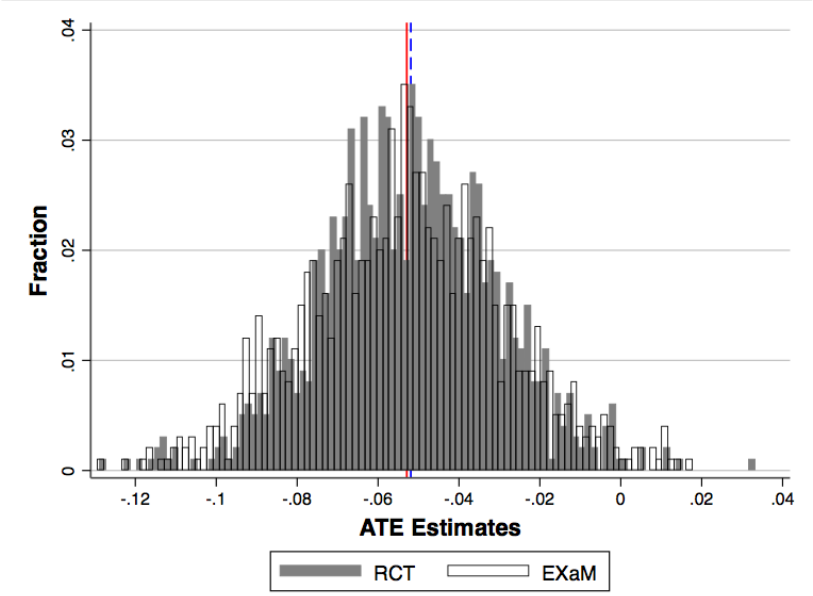
(b) Avg Predicted Effects of Assigned Treatments e_i^*



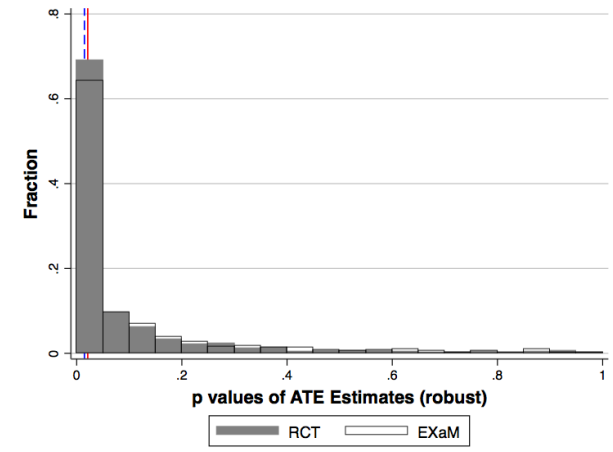
Notes: To compare EXaM and RCT's welfare performance, this figure shows the distribution of average subject welfare over 1000 bootstrap simulations under each experimental design. Panel 2a measures welfare with respect to average WTP w_i^* of assigned treatments while Panel 2b with respect to average predicted effects e_i^* of assigned treatments. A dotted line indicates the distribution of each welfare measure for RCT while a solid line indicates that for EXaM. Each vertical line represents median. I simulate values of WTP $\hat{w}_{it_1}$ and predicted effects $\hat{e}_{t_1 i}$ with the main statistical specification including all of the interactions between the treatment indicator and household characteristics (baseline latrine density, diarrhea prevention knowledge score, mother's years of education, and having a boy in the household). See Section 7.3 for discussions about this figure. See Appendix A.2.4 for the detailed computational procedure.

Figure 3: EXaM vs RCT: Information

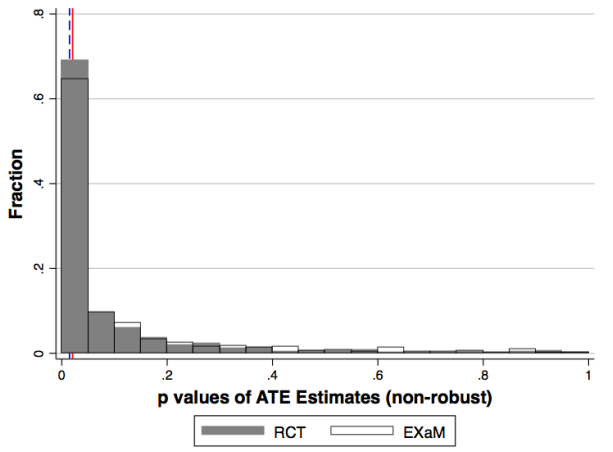
(a) Distribution of Average Treatment Effect Estimates



(b) Associated p Values (Robust)

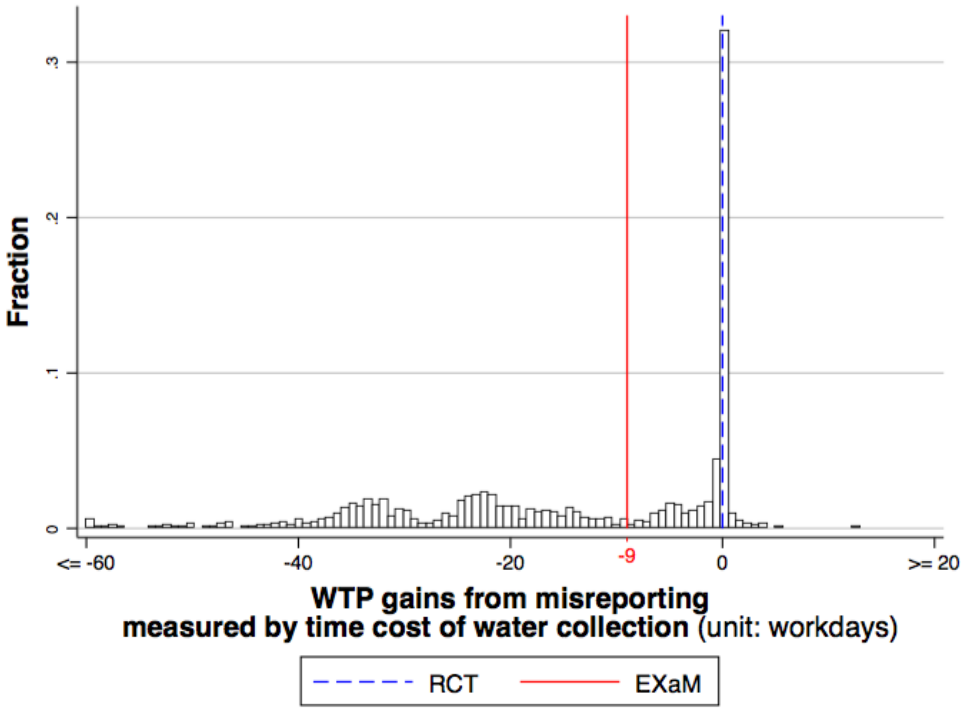


(c) Associated p Values (Non-robust)



Notes: This figure compares EXaM and RCT's causal inference performance by showing the randomization distribution of average treatment effect estimates and accompanying p values under each experimental design. Grey bins indicate average treatment effect estimates for RCT while transparent bins with black outlines indicate those for EXaM. The solid vertical line indicates median for EXaM while the dashed vertical line indicates that for RCT. See Section 7.3 for discussions about this figure. See Appendix A.2.4 for the detailed computational procedure.

Figure 4: EXaM vs RCT: Incentive



Notes: To quantify the incentive compatibility of EXaM, this figure shows the histogram of true WTP gains from potential WTP misreports to EXaM. The solid vertical line represents median. The dash vertical line is for RCT, where the true WTP gain from any WTP misreport is zero. See Section 7.3 for discussions about this figure. See Appendix A.2.4 for the detailed computational procedure.

A Appendix

A.1 Methodological Appendix

A.1.1 Proposition 5: Generalization

This section extends Proposition 5 to a general case where $p_t n_p$ may not be an integer. Let $N_{pt} \equiv \sum_i 1\{p_i^*(\epsilon) = p\} D_{it}$ be a random variable that stands for the number of subjects with propensity vector p and assigned to treatment t. Denote the realization of N_{pt} by $n_{pt} \equiv \sum_i 1\{p_i^*(\epsilon) = p\} d_{it}$. Let $\underline{n}_{pt}$ be the greatest integer less than or equal to $p_t n_p$, the expected number of subjects with propensity vector p and assigned to treatment t. I first extend Definition 2 as follows.

Definition 2 (EXaM Continued; Generalization).

(3) Draw a treatment assignment from p_{it} as follows. I first use Budish et al. (2013)'s network-flow algorithm (in their Appendix B) to draw (n_{pt}) that satisfy the following properties:[38]

- $n_{pt} = \underline{n}_{pt}$ for all p and t such that $p_t n_p \in \mathbb{N}$.
- $n_{pt} \in \{\underline{n}_{pt}, \underline{n}_{pt} + 1\}$ for all p and t such that $p_t n_p \notin \mathbb{N}$.
- $\sum_t n_{pt} = n_p$ for all p.
- $\sum_p n_{pt} = c_t$ for all t.

Given the drawn values of (n_{pt}), for each propensity vector p,

- Step 1: I uniformly randomly pick $p_{t_0} n_p$ subjects from $\{i | p_i^*(\epsilon) = p\}$ and assign them to t_0.

[38] To do so, I embed my setting into their notation as follows:

- $N = \{p | \text{there exists some subject } i \text{ such that } p_i^*(\epsilon) = p\}$.
- $O = \{t_0, t_1, ..., t_m\}$.
- $\mathcal{H} = \{\mathcal{H}_0, \mathcal{H}_1, \mathcal{H}_2\}$ where $\mathcal{H}_0 = \{(p,t) | p \in N, t \in O\}, \mathcal{H}_1 = \{(p,t) | t \in O\}_{p \in N}$, and $\mathcal{H}_2 = \{(p,t) | p \in N\}_{t \in O}$.
- $\bar{q}_s = 1$ and $\underline{q}_s = 0$ if $s \in \mathcal{H}_0$.
- $\bar{q}_s = \underline{q}_s = n_p - \Sigma_t \underline{n}_{pt}$ if $s \in \mathcal{H}_1$.
- $\bar{q}_s = \underline{q}_s = c_t - \Sigma_p \underline{n}_{pt}$ if $s \in \mathcal{H}_2$.

The above specification of $\mathcal{H}$ satisfies the "bihierarchy" condition in Budish et al. (2013)'s Theorem 1. Their Theorem 1 and Appendix B therefore imply that the output of their network-flow algorithm for this problem satisfies the properties above.

For each subsequent step $k = 1, ..., m$,

- Step k: From the remaining $n_p - \Sigma_{t=t_0}^{t_{k-1}} p_t n_p$ subjects, I uniformly randomly pick $p_{t_k} n_p$ subjects and assign them to t_k.

For this general setting and extended Definition 2, I obtain the following characterization of the variance of $\hat{\beta}_t$, which nests Proposition 5 in Section 5.

Proposition 5 (Generalization).

$$V(\hat{\beta}_t|p^*(\epsilon)) = \sum_p \delta_p^2 \Big\{ \sum_{t' \in \{t_0, t\}} \Big[\Big(\frac{S_{pt'}^2}{\underline{n}_{pt'}} \Big)(1 - p_{t'} n_p + \underline{n}_{pt'}) + \Big(\frac{S_{pt'}^2}{\underline{n}_{pt'} + 1} \Big)(p_{t'} n_p - \underline{n}_{pt'}) \Big] - \frac{S_{ptt_0}^2}{n_p} \Big\}. \tag{12}$$

A.1.2 Proofs

Proof of Proposition 1

Suppose to the contrary that there exist some $\epsilon \in [0, \bar{\epsilon}), i$, and t such that $p_{it}^*(\epsilon) \neq p_{it}^{RCT}$. Since $e_{ti} = e_{t'j}$ for all subjects i and j and treatments t and t', I have $\pi_{te_{ti}} \equiv \alpha e_{ti} + \beta_t = \alpha e_{tj} + \beta_t \equiv \pi_{te_{tj}}$ for all subjects i and j and treatment t. Combined with $w_{it} = w_{jt'}$ for all subjects i and j and treatments t and t', this implies that any subjects i and j face the same utility maximization problem:

$$\arg\max_{p_i \in P}(\Sigma_t p_{it} w_{it} \text{ s.t. } \Sigma_t p_{it} \pi_{te_{ti}} \leq b) = \arg\max_{p_j \in P}(\Sigma_t p_{jt} w_{jt} \text{ s.t. } \Sigma_t p_{jt} \pi_{te_{tj}} \leq b).$$

This implies $p_{jt}^*(\epsilon) = p_{it}^*(\epsilon) \neq p_{it}^{RCT} \equiv c_t/n$ by the assumption that all subjects use the common tie-breaking rule (uniformly mixing among cheapest utility-maximizing p_i's). If $p_{jt}^*(\epsilon) = p_{it}^*(\epsilon) > c_t/n$, then $\Sigma_{j=1}^n p_{jt}^*(\epsilon) = n p_{it}^*(\epsilon) > n c_t/n = c_t$, contradicting the capacity constraint in the definition of $p_{it}^*(\epsilon)$. If $p_{jt}^*(\epsilon) = p_{it}^*(\epsilon) < c_t/n$, then there is another treatment t' for which $p_{jt'}^*(\epsilon) = p_{it'}^*(\epsilon) > c_t/n$ since $\Sigma_t c_t/n = \Sigma_t p_{jt}^*(\epsilon) = 1$ for any subject j. This implies that $\Sigma_{j=1}^n p_{jt'}^*(\epsilon) = n p_{it'}^*(\epsilon) > n c_{t'}/n = c_{t'}$, again contradicting the capacity constraint. Thus, for every $\epsilon \in [0, \bar{\epsilon}), i$, and t, it must be the case that $p_{it}^*(\epsilon) = p_{it}^{RCT}$.

Proof of Proposition 2

EXaM always exists: For every treatment t, fix α_t at any negative constant $\alpha^* < 0$. Define a space of possible values of $\beta \equiv (\beta_t)_t$ by $B \equiv [0, nb - \alpha^* \bar{e} 1\{\bar{e} > 0\}]^{m+1}$ where $\bar{e} = \max\{e_{ti}\}$. Define the demand correspondence for each subject i by $p_i^*(\beta) \equiv \text{argmax}_{p_i \in P}(\Sigma_t p_{it} w_{it}$ s.t.

$\Sigma_t p_{it}(\alpha^* e_{ti} + \beta_t) \leq b$). Define the excess demand correspondence $z(\cdot) : B \to \mathbb{R}^{m+1}$ by $z(\beta) = \{\Sigma_i p_i - c | p_i \in p_i^*(\beta) \text{ for every } i\} \equiv \Sigma_i p_i^*(\beta) - c$ where $c \equiv (c_t)$. This correspondence $z(\cdot)$ is upper hemicontinuous and convex-valued because it is a linear finite sum of $p_i^*(\beta)$'s, which are upper hemicontinuous and convex-valued as shown below.

Lemma 1. *For every subject i, her demand correspondence $p_i^*(\beta)$ is nonempty, convex-valued for every $\beta \in B$, and upper hemicontinuous in β.*

Proof of Lemma 1. $p_i^*(\beta)$ is convex-valued since the utility function is linear and for any $p_i, p_i' \in p_i^*(\beta) \subset P$ and $\delta \in [0,1]$, it holds that $\delta p_i + (1-\delta)p_i'$ is in P (since P is convex) and satisfies the budget constraint (since $\Sigma_t[\delta p_{it} + (1-\delta)p_{it}'](\alpha^* e_{ti} + \beta_t) = \delta \Sigma_t p_{it}(\alpha^* e_{ti} + \beta_t) + (1-\delta)\Sigma_t p_{it}'(\alpha^* e_{ti} + \beta_t) \leq \delta b + (1-\delta)b = b$). $p_i^*(\beta)$ is non-empty and upper-hemicontinuous by the maximum theorem. To see this, note that (1) the utility function is linear and (2) the correspondence from β to the choice set $\{p_i \in P | \Sigma_t p_{it}(\alpha^* e_{ti} + \beta_t) \leq b\}$ is both upper-hemicontinuous and lower-hemicontinuous as well as compact-valued. Thus the maximum theorem implies that $p_i^*(\beta)$ is non-empty and upper-hemicontinuous, completing the proof of Lemma 1. □

Let $\bar{c} \equiv \max_{t=t_0,t_1,\ldots,t_m} c_t$ and $\tilde{B} = [-\bar{c}, n(b+\bar{c}) - \alpha^* \bar{e} 1\{\bar{e} > 0\}]^{m+1}$. Define a truncation function $f : \tilde{B} \to B$ by $f(\beta) \equiv (\max\{0, \min\{\beta_t, nb - \alpha^* \bar{e} 1\{\bar{e} > 0\}\}\})_{t=t_0,t_1,\ldots,t_m}$. Define correspondence $g : \tilde{B} \to B$ by $g(\beta) \equiv f(\beta) + z(f(\beta))$.

Lemma 2. *g has a fixed point $\beta^* \in g(\beta^*)$.*

Proof of Lemma 2. $z(f(\beta))$ is upper hemicontinuous and convex-valued as a function of $\beta \in \tilde{B}$ because $f(\cdot)$ is continuous and $z(\cdot)$ is an upper hemicontinuous and convex-valued correspondence, as explained above. This implies that $g(\beta)$ is upper hemicontinuous and convex-valued as well. The range of $g(\beta)$ lies in $\tilde{B}$, i.e., $g : \tilde{B} \to \tilde{B}$. It is because

- $f(\beta) \equiv (\max\{0, \min\{\beta_t, nb - \alpha^* \bar{e} 1\{\bar{e} > 0\}\}\})_{t=t_0,t_1,\ldots,t_m} \in [0, nb - \alpha^* \bar{e} 1\{\bar{e} > 0\}]^{m+1}$, which is by $nb - \alpha^* \bar{e} 1\{\bar{e} > 0\} \geq 0$
- $\bar{c} \equiv \max_{t=t_0,t_1,\ldots,t_m} c_t \geq 1$
- $z(f(\beta)) \in [-\bar{c}, n]^{m+1}$ because, for any $\beta \in \tilde{B}$ and t, the excess demand $z_t(\beta)$ is at least $-\bar{c}$ (since the supply of any treatment t is $c_t \leq \bar{c}$ by definition) and at most n (since there are n subjects the demand of any treatment t by any subject i is at most 1)

Finally, $\tilde{B}$ is nonempty by $-\bar{c} < 0 < n(b+\bar{c}) \leq n(b+\bar{c}) - \alpha^* \bar{e} 1\{\bar{e} > 0\}$. $g(\beta) \equiv f(\beta) + z(f(\beta))$ is therefore an upper hemicontinuous, nonempty, and convex-valued correspondence defined

on the non-empty, compact, and convex set $\tilde{B}$. By Kakutani's fixed point theorem, there exists a fixed point $\beta^* \in g(\beta^*)$, proving Lemma 2. □

Lemma 3. *For any fixed point β^* of $g(\cdot)$, the associated price function vector $(\pi_{te} \equiv \alpha^* e + f_t(\beta^*))_t$, where $f_t(\beta^*)$ is the t-th element of $f(\beta^*)$, satisfies the conditions of EXaM.*

Proof of Lemma 3. By the definition of a fixed point and correspondence $g(\cdot)$, there exists $z^* \equiv (z_t^*) \in z(f(\beta^*))$ such that $\beta_t^* = f_t(\beta^*) + z_t^*$ for all t. Fix any such z^* and the associated β^*. It is enough to show that the associated equilibrium treatment assignment probability vector $(p_{it}^*)_{i,t} \in \text{argmax}_{p_i \in P}(\Sigma_t p_{it} w_{it} \text{ s.t. } \Sigma_t p_{it}(\alpha^* e_{ti} + f_t(\beta^*)) \leq b)$ satisfies the capacity constraint for every treatment t. For each treatment t, there are three cases to consider:

Case 1: $\beta_t^* < 0$. Then $f_t(\beta^*) \equiv \max\{0, \min\{\beta_t^*, nb - \alpha^* \bar{e} 1\{\bar{e} > 0\}\}\} = 0$ and hence $\beta_t^* = f_t(\beta^*) + z_t^*$ implies $\beta_t^* = z_t^* \equiv \sum_i p_{it}^* - c_t < 0$, implying $\sum_i p_{it}^* < c_t$, i.e., the capacity constraint holds.

Case 2: $\beta_t^* \in [0, nb - \alpha^* \bar{e} 1\{\bar{e} > 0\}]$. By the definition of f, I have $f_t(\beta^*) = \beta_t^*$. Then $\beta_t^* = f_t(\beta^*) + z_t^*$ implies $z_t^* = 0$, i.e., the capacity constraint holds with equality.

Case 3: $\beta_t^* > nb - \alpha^* \bar{e} 1\{\bar{e} > 0\}$. Then $f_t(\beta^*) = nb - \alpha^* \bar{e} 1\{\bar{e} > 0\}$ and hence $\beta_t^* = f_t(\beta^*) + z_t^*$ implies that $z_t^* = \beta_t^* - nb - \alpha^* \bar{e} 1\{\bar{e} > 0\} > 0$, i.e., treatment t is in excess demand at price $\pi_{te} \equiv \alpha^* e + f_t(\beta^*)$. However, for any possible predicted effect level $e \leq \bar{e}$, we have

$$\pi_{te} \equiv \alpha^* e + f_t(\beta^*) = \alpha^* e + nb - \alpha^* \bar{e} 1\{\bar{e} > 0\} = \begin{cases} nb + \alpha^*(e - \bar{e}) \geq nb & \text{if } \bar{e} > 0 \\ nb + \alpha^* \bar{e} \geq nb & \text{otherwise,} \end{cases}$$

where the last inequality is by $\alpha^* < 0$ and $e \leq \bar{e} \leq 0$. Therefore, for each subject i, $p_{it}^* \leq b / \pi_{te_{ti}} \leq 1/n$. This implies that $\Sigma_i p_{it}^* \leq 1 \leq c_t$. This completes the proof of Lemma 3. □

EXaM is ex ante Pareto efficient subject to the randomization constraint: Suppose to the contrary that there exists $\epsilon \in [0, \bar{\epsilon})$ such that $p_{it}^*(\epsilon)$ is ex ante Pareto dominated by another feasible treatment assignment probabilities $(p_{it}(\epsilon))_{i,t} \in P^n$ with $p_{it}(\epsilon) \in [\epsilon, 1 - \epsilon]$ for all i and t, i.e.,

- $\Sigma_t p_{it}(\epsilon) e_{ti} \geq \Sigma_t p_{it}^*(\epsilon) e_{ti}$ for all i and
- $\Sigma_t p_{it}(\epsilon) w_{it} \geq \Sigma_t p_{it}^*(\epsilon) w_{it}$ for all i

with at least one strict inequality. Let me use $p_{it}(\epsilon)$ to define the following treatment assignment probabilities:

$$p_{it} \equiv [p_{it}(\epsilon) - qp_{it}^{RCT}]/(1-q)$$

where $q \equiv \inf\{q' \in [0,1] | (1-q')p_{it}^* + q'p_{it}^{RCT} \in [\epsilon, 1-\epsilon] \text{ for all } i \text{ and } t\}$ is the mixing weight used for defining and computing $p_{it}^*(\epsilon)$ in Definition 2. In other words, p_{it} are the treatment assignment probabilities such that the following holds:

$$p_{it}(\epsilon) \equiv (1-q)p_{it} + qp_{it}^{RCT}.$$

Since both $p_{it}(\epsilon)$ and p_{it}^{RCT} are feasible treatment assignment probabilities in P^n, p_{it} is also feasible and in P^n (note that $\Sigma_t p_{it} = \Sigma_t[p_{it}(\epsilon) - qp_{it}^{RCT}]/(1-q)) = (1-q)/(1-q) = 1$ for every i). For each i, I have

$$\begin{aligned}
&\Sigma_t p_{it}(\epsilon)e_{ti} \geq \Sigma_t p_{it}^*(\epsilon)e_{ti} \\
&\Leftrightarrow \Sigma_t((1-q)p_{it} + qp_{it}^{RCT})e_{ti} \geq \Sigma_t((1-q)p_{it}^* + qp_{it}^{RCT})e_{ti} \\
&\Leftrightarrow \Sigma_t(1-q)p_{it}e_{ti} \geq \Sigma_t(1-q)p_{it}^*e_{ti} \\
&\Leftrightarrow \Sigma_t p_{it}e_{ti} \geq \Sigma_t p_{it}^*e_{ti}.
\end{aligned}$$

Similarly, for each i, I have

$$\begin{aligned}
&\Sigma_t p_{it}(\epsilon)w_{it} \geq \Sigma_t p_{it}^*(\epsilon)w_{it} \\
&\Leftrightarrow \Sigma_t((1-q)p_{it} + qp_{it}^{RCT})w_{it} \geq \Sigma_t((1-q)p_{it}^* + qp_{it}^{RCT})w_{it} \\
&\Leftrightarrow \Sigma_t(1-q)p_{it}w_{it} \geq \Sigma_t(1-q)p_{it}^*w_{it} \\
&\Leftrightarrow \Sigma_t p_{it}w_{it} \geq \Sigma_t p_{it}^*w_{it}.
\end{aligned}$$

Therefore, the assumption that $p_{it}(\epsilon)$ ex ante Pareto dominates $p_{it}^*(\epsilon)$ implies that p_{it} ex ante Pareto dominates p_{it}^*, i.e.,

- $\Sigma_t p_{it}e_{ti} \geq \Sigma_t p_{it}^*e_{ti}$ for all i and
- $\Sigma_t p_{it}w_{it} \geq \Sigma_t p_{it}^*w_{it}$ for all i

with at least one strict inequality. There are two cases to consider.

Case 1: $\Sigma_t p_{\tilde{i}t}e_{t\tilde{i}} > \Sigma_t p_{it}^*e_{t\tilde{i}}$ for some $\tilde{i}$. This implies

$$\Sigma_t \Sigma_i p_{it}e_{ti} > \Sigma_t \Sigma_i p_{it}^*e_{ti}$$

$\Leftrightarrow \Sigma_t \Sigma_i p_{it}(\pi_{te_{ti}} - \beta_t)/\alpha > \Sigma_t \Sigma_i p^*_{it}(\pi_{te_{ti}} - \beta_t)/\alpha$
(by the definition of $\pi_{te} \equiv \alpha e + \beta_t$ with $\alpha \neq 0$)
$\Leftrightarrow \Sigma_t \Sigma_i p_{it} \pi_{te_{ti}}/\alpha > \Sigma_t \Sigma_i p^*_{it} \pi_{te_{ti}}/\alpha$
(since $\Sigma_i p_{it} = \Sigma_i p^*_{it} = c_t$)
$\Leftrightarrow \Sigma_t \Sigma_i p_{it} \pi_{te_{ti}} < \Sigma_t \Sigma_i p^*_{it} \pi_{te_{ti}}$.
(since $\alpha < 0$ by Definition 2)

I thus have

$$\Sigma_t \Sigma_i p_{it} \pi_{te_{ti}} < \Sigma_t \Sigma_i p^*_{it} \pi_{te_{ti}}. \tag{13}$$

However, it has also to be the case that $\Sigma_t p_{it} \pi_{te_{ti}} \geq \Sigma_t p^*_{it} \pi_{te_{ti}}$ for any i since (a) $\Sigma_t p_{it} w_{it} \geq \Sigma_t p^*_{it} w_{it}$ by assumption and (b) $(p^*_{it})_t$ is (a mixture of) the cheapest among all feasible assignment probability vectors that i most prefers under prices $(\pi_{te})_{t,e}$ and budget b. Thus $\Sigma_t \Sigma_i p_{it} \pi_{te_{ti}} \geq \Sigma_t \Sigma_i p^*_{it} \pi_{te_{ti}}$, a contradiction to inequality (13).

Case 2: $\Sigma_t p_{\tilde{i}t} w_{\tilde{i}t} > \Sigma_t p^*_{it} w_{\tilde{i}t}$ for some $\tilde{i}$. Since $\tilde{i}$ most prefers $(p^*_{\tilde{i}t})_t$ among all feasible assignment probability vectors that satisfies the budget constraint under prices $(\pi_{te})_{t,e}$, the strictly more preferred treatment assignment probability vector $(p_{\tilde{i}t})_t$ must violate the budget constraint, i.e., $\Sigma_t p_{\tilde{i}t} \pi_{te_{t\tilde{i}}} > b \geq \Sigma_t p^*_{\tilde{i}t} \pi_{te_{t\tilde{i}}}$, where the second weak inequality comes from the assumption that $(p^*_{\tilde{i}t})_t$ satisfies the budget constraint under prices (π_{te}). Moreover, for any other subject $i \neq \tilde{i}$, $\Sigma_t p_{it} \pi_{te_{ti}} \geq \Sigma_t p^*_{it} \pi_{te_{ti}}$ since $(p^*_{it})_t$ is (a mixture of) the cheapest among all assignment probability vectors in P^n that i most prefers under prices $(\pi_{te})_{t,e}$ and budget b. I thus have

$$\begin{aligned} &\Sigma_t p_{\tilde{i}t} \pi_{te_{t\tilde{i}}} + \Sigma_{i \neq \tilde{i}} \Sigma_t p_{it} \pi_{te_{ti}} > \Sigma_t p^*_{\tilde{i}t} \pi_{te_{t\tilde{i}}} + \Sigma_{i \neq \tilde{i}} \Sigma_t p^*_{it} \pi_{te_{ti}} \\ \Leftrightarrow\ &\Sigma_i \Sigma_t p_{it} \pi_{te_{ti}} > \Sigma_i \Sigma_t p^*_{it} \pi_{te_{ti}}. \end{aligned}$$

However, by the logic described in Case 1, the assumption ($\Sigma_i p_{it} e_{ti} \geq \Sigma_i p^*_{it} e_{ti}$ for all t) implies that $\Sigma_i \Sigma_t p_{it} \pi_{te_{ti}} \leq \Sigma_i \Sigma_t p^*_{it} \pi_{te_{ti}}$, a contradiction.

Therefore, $p^*_{it}(\epsilon)$ with any $\epsilon \in [0, \bar{\epsilon})$ is never ex ante Pareto dominated by another feasible treatment assignment probabilities $(p_{it}(\epsilon))_{i,t} \in P^n$ with $p_{it}(\epsilon) \in [\epsilon, 1-\epsilon]$ for all i and t.

Proof of Proposition 3

The proof uses intermediate observations.

Lemma 4. *EXaM is "envy-free," i.e., for any experimental design problem, any $\epsilon \in [0, \bar{\epsilon}]$, any subjects i and j with $e_{ti} = e_{tj}$ for all t,*

$$\Sigma_t p^*_{it}(\epsilon) w_{it} \geq \Sigma_t p^*_{jt}(\epsilon) w_{it}$$

Proof of Lemma 4. In Definition 2, all subjects have the same budget and any subjects i and j with $e_{ti} = e_{tj}$ face the same price π_{te} of treatment t. For any subjects i and j with $e_{ti} = e_{tj}$ for all t, therefore, $(p^*_{jt})_t$ satisfies i's budget constraint and $\Sigma_t p^*_{it} w_{it} \geq \Sigma_t p^*_{jt} w_{it}$. This implies the desired conclusion since

$$\begin{aligned}
&\Sigma_t p^*_{it} w_{it} \geq \Sigma_t p^*_{jt} w_{it} \\
\Leftrightarrow\ &(1-q)\Sigma_t p^*_{it} w_{it} + q\Sigma_t p^{RCT}_{it} w_{it} \geq (1-q)\Sigma_t p^*_{jt} w_{it} + q\Sigma_t p^{RCT}_{jt} w_{it} \\
\Leftrightarrow\ &\Sigma_t p^*_{it}(\epsilon) w_{it} \geq \Sigma_t p^*_{jt}(\epsilon) w_{it},
\end{aligned}$$

where the first equivalence is by $p^{RCT}_{it} = p^{RCT}_{jt} \equiv c_t/n$. □

Lemma 5. *EXaM with WTP reporting is "semi-anonymous." That is, for any sequence of experimental design problems, any n with any $\epsilon^n \in [0, \bar{\epsilon}]$, any subjects i and j with $e_{ti} = e_{tj}$ for all t, let $(w_i, w_j, w_{-\{i,j\}})$ be a permutation of $(w_j, w_i, w_{-\{i,j\}})$ obtained by permuting i and j's WTP reports w_i and w_j. Semi-anonymity means that*

$$p^{*n}_i(w_i, w_j, w_{-\{i,j\}}; \epsilon^n) = p^{*n}_j(w_j, w_i, w_{-\{i,j\}}; \epsilon^n) \text{ and}$$
$$p^{*n}_j(w_i, w_j, w_{-\{i,j\}}; \epsilon^n) = p^{*n}_i(w_j, w_i, w_{-\{i,j\}}; \epsilon^n)$$

Proof of Lemma 5. In Definition 2 of EXaM, all subjects have the same budget and any subjects i and j with $e_{ti} = e_{tj}$ face the same price π_{te} of treatment t. For any subjects i and j with $e_{ti} = e_{tj}$ for all t, therefore, given any $w_{-\{i,j\}}$, subject i with WTP report w_j solves the same constrained utility maximization problem as subject j with WTP report w_j does. Therefore, $p^{*n}_i(w_i, w_j, w_{-\{i,j\}}; 0) = p^{*n}_j(w_j, w_i, w_{-\{i,j\}}; 0)$ and $p^{*n}_j(w_i, w_j, w_{-\{i,j\}}; 0) = p^{*n}_i(w_j, w_i, w_{-\{i,j\}}; 0)$. This implies semi-anonymity since

$$\begin{aligned}
&p^{*n}_i(w_i, w_j, w_{-\{i,j\}}; \epsilon^n) \\
\equiv\ &(1-q^n) p^{*n}_i(w_i, w_j, w_{-\{i,j\}}; 0) + q^n p^{RCTn}_i \\
=\ &(1-q^n) p^{*n}_j(w_j, w_i, w_{-\{i,j\}}; 0) + q^n p^{RCTn}_j \\
\equiv\ &p^{*n}_j(w_j, w_i, w_{-\{i,j\}}; \epsilon^n),
\end{aligned}$$

where q^n is the mixing probability q for the n-th problem in the sequence of experimental design problems while $p_i^{RCTn} = p_j^{RCTn} \equiv c_t^n/n$. □

Lemmas 4 and 5 imply Proposition 3 by using Theorem 1 of Azevedo and Budish (2017) (precisely, a generalization of their Theorem 1 in their Supplementary Appendix B).

A Statistical Lemma

Lemma 6. *Assume a sample of m subjects is randomly drawn (i.e., every combination of m subjects occurs with equal probability) from the fixed finite population of n subjects with a fixed vector of a variable $(X_1, ...X_n)$. Denote the random sample by I. Let $\mu = \frac{1}{n}\sum_{i=1}^n X_i$, $\sigma^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \mu)^2$, $\hat{\mu} = \frac{1}{m}\sum_{i\in I} X_i$ and $\hat{\sigma}^2 = \frac{1}{m-1}\sum_{i\in I}(X_i - \hat{\mu})^2$. Then,*

$$V(\hat{\mu}) = \frac{n-m}{nm}\sigma^2 \text{ and } E(\hat{\sigma}^2) = \sigma^2.$$

Proof of Lemma 6. Let $W_i = 1\{i \in I\}$ so that $\hat{\mu} = \frac{1}{m}\sum_{i=1}^n X_i W_i$ and $\hat{\sigma}^2 = \frac{1}{m-1}\sum_{i=1}^n (X_i - \hat{\mu})^2 W_i$. Then $E(W_i) = E(W_i^2) = \frac{m}{n}$ for all i, implying $V(W_i) = \frac{m}{n} - (\frac{m}{n})^2 = \frac{m(n-m)}{n^2}$ for all i. Since $E(W_i W_j) = \frac{m(m-1)}{n(n-1)}$ for any $i \neq j$, it is the case that for any $i \neq j$,

$$\begin{aligned}
&Cov(W_i, W_j) \\
&= E[(W_i - \frac{m}{n})(W_j - \frac{m}{n})] \\
&= E(W_i W_j) - \frac{m}{n}E(W_j) - \frac{m}{n}E(W_i) + (\frac{m}{n})^2 \\
&= \frac{m(m-1)}{n(n-1)} - (\frac{m}{n})^2 \\
&= -\frac{m(n-m)}{n^2(n-1)}
\end{aligned}$$

It follows that

$$\begin{aligned}
&V(\hat{\mu}) \\
&= \frac{1}{m^2}V(\textstyle\sum_{i=1}^n X_i W_i) \\
&= \frac{1}{m^2}(\textstyle\sum_{i=1}^n X_i^2 V(W_i) + \sum_{i=1}^n \sum_{j\neq i} X_i X_j Cov(W_i, W_j)) \\
&= \frac{1}{m^2}(\frac{m(n-m)}{n^2}\textstyle\sum_{i=1}^n X_i^2 - \frac{m(n-m)}{n^2(n-1)}\sum_{i=1}^n \sum_{j\neq i} X_i X_j)
\end{aligned}$$

$$= \frac{n-m}{n^2 m}(\sum_{i=1}^n X_i^2 - \frac{1}{n-1}\sum_{i=1}^n \sum_{j\neq i} X_i X_j)$$

$$= \frac{n-m}{n^2 m}(\sum_{i=1}^n X_i^2 - \frac{1}{n-1}\sum_{i=1}^n \sum_{j=1}^n X_i X_j + \frac{1}{n-1}\sum_{i=1}^n X_i^2)$$

$$= \frac{n-m}{n^2 m}(\frac{n}{n-1}\sum_{i=1}^n X_i^2 - \frac{1}{n-1}\sum_{i=1}^n \sum_{j=1}^n X_i X_j)$$

$$= \frac{n-m}{nm(n-1)}(\sum_{i=1}^n X_i^2 - \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n X_i X_j)$$

$$= \frac{n-m}{nm(n-1)}(\sum_{i=1}^n X_i^2 - \frac{2}{n}\sum_{i=1}^n X_i \sum_{j=1}^n X_j + \frac{1}{n}(\sum_{i=1}^n X_i)^2)$$

$$= \frac{n-m}{nm(n-1)}\sum_{i=1}^n (X_i^2 - \frac{2}{n} X_i \sum_{j=1}^n X_j + \frac{1}{n^2}(\sum_{j=1}^n X_j)^2)$$

$$= \frac{n-m}{nm(n-1)}\sum_{i=1}^n (X_i - \frac{1}{n}\sum_{j=1}^n X_j)^2$$

$$= \frac{n-m}{nm}\sigma^2.$$

For the other part,

$$E(\hat{\sigma}^2)$$

$$= \frac{1}{m-1}E(\sum_{i=1}^n (X_i - \hat{\mu})^2 W_i)$$

$$= \frac{1}{m-1}E(\sum_{i=1}^n (X_i^2 W_i - 2X_i W_i \hat{\mu} + \hat{\mu}^2 W_i))$$

$$= \frac{1}{m-1}E(\sum_{i=1}^n X_i^2 W_i - 2m\hat{\mu}^2 + m\hat{\mu}^2)$$

$$= \frac{1}{m-1}(\frac{m}{n}\sum_{i=1}^n X_i^2 - m(V(\hat{\mu}) + [E(\hat{\mu})]^2))$$

$$= \frac{1}{m-1}(\frac{m}{n}\sum_{i=1}^n X_i^2 - \frac{n-m}{n}\sigma^2 - m\mu^2)$$

$$= \frac{1}{m-1}(\frac{m(n-1)}{n}\sigma^2 - \frac{n-m}{n}\sigma^2)$$

$$= \sigma^2.$$ □

Proof of Proposition 4

The proof uses the following lemma.

Lemma 7. *There exists estimator $\hat{\theta}_{EXaM,t}$ such that $E(\hat{\theta}_{EXaM,t}|p^*(\epsilon)) = E(\hat{\mu}(t)^2|p^{RCT})$.*

Proof of Lemma 7. Let $\hat{\mu}_{RCT,t} = \hat{\mu}(t)$, $\mu_t = \frac{1}{n}\sum_{i=1}^n Y_i(t)$ and $S_t^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i(t) - \mu_t)^2$. We have

$$\begin{aligned}
E(\hat{\mu}_{RCT,t}^2|p^{RCT}) &= Var(\hat{\mu}_{RCT,t}|p^{RCT}) + E(\hat{\mu}_{RCT,t}|p^{RCT})^2 \\
&= \frac{n-c_t}{nc_t}S_t^2 + \mu_t^2 \\
&= \frac{n-c_t}{nc_t}\frac{1}{n-1}(\sum_{i=1}^n Y_i(t)^2 - n\mu_t^2) + \mu_t^2 \\
&= \frac{n-c_t}{(n-1)c_t}\frac{1}{n}\sum_{i=1}^n Y_i(t)^2 + \frac{n(c_t-1)}{(n-1)c_t}\mu_t^2, \qquad (14)
\end{aligned}$$

where the second equality holds by the first part of Lemma 6 and the fact that $E(\hat{\mu}_{RCT,t}|p^{RCT}) = \mu_t$. Under EXaM $p^*(\epsilon)$, $\hat{\theta}_{1t} = \frac{1}{n}\sum_p \frac{1}{p_t}\sum_{i:p_i^*(\epsilon)=p} Y_i^2 D_{it}$ unbiasedly estimates $\frac{1}{n}\sum_{i=1}^n Y_i(t)^2$ because

$$\begin{aligned}
E(\hat{\theta}_{1t}|p^*(\epsilon)) &= \frac{1}{n}\sum_p \frac{1}{p_t}\sum_{i:p_i^*(\epsilon)=p} Y_i(t)^2 E(D_{it}|p^*(\epsilon)) \\
&= \frac{1}{n}\sum_p \frac{1}{p_t}\sum_{i:p_i^*(\epsilon)=p} Y_i(t)^2 p_t \\
&= \frac{1}{n}\sum_p \sum_{i:p_i^*(\epsilon)=p} Y_i(t)^2 \\
&= \frac{1}{n}\sum_{i=1}^n Y_i(t)^2.
\end{aligned}$$

Next I obtain an unbiased estimator for μ_t^2 under EXaM $p^*(\epsilon)$. Let $\mu_{pt} = \frac{1}{n_p}\sum_{i:p_i^*(\epsilon)=p} Y_i(t)$, $\hat{\mu}_{EXaM,pt} = \frac{1}{p_t n_p}\sum_{i:p_i^*(\epsilon)=p} Y_i D_{it}$ and $\hat{\mu}_{EXaM,t} = \sum_p \frac{n_p}{n}\hat{\mu}_{EXaM,pt}$. Note that $E(\hat{\mu}_{EXaM,pt}|p^*(\epsilon)) = \mu_{pt}$ and $\mu_t = \sum_p \frac{n_p}{n}\mu_{pt}$. Note also that, by Definition 2 (3), treatment assignments are independent across subpopulations with different propensities. Hence, $\hat{\mu}_{EXaM,pt}$ is independent across p. With $S_{pt}^2 = \frac{1}{n_p - 1}\sum_{i:p_i^*(\epsilon)=p}(Y_i(t) - \mu_{pt})^2$, I have

$$
\begin{aligned}
E(\hat{\mu}^2_{EXaM,t}|p^*(\epsilon)) &= E((\sum_p \frac{n_p}{n}\hat{\mu}_{EXaM,pt})^2|p^*(\epsilon)) \\
&= \sum_p \sum_{p' \neq p} \frac{n_p n_{p'}}{n^2} E(\hat{\mu}_{EXaM,pt}|p^*(\epsilon)) E(\hat{\mu}_{EXaM,p't}|p^*(\epsilon)) + \sum_p \frac{n_p^2}{n^2} E(\hat{\mu}^2_{EXaM,pt}|p^*(\epsilon)) \\
&= \sum_p \sum_{p' \neq p} \frac{n_p n_{p'}}{n^2} \mu_{pt}\mu_{p't} + \sum_p \frac{n_p^2}{n^2}(Var(\hat{\mu}_{EXaM,pt}|p^*(\epsilon)) + E(\hat{\mu}_{EXaM,pt}|p^*(\epsilon))^2) \\
&= \sum_p \sum_{p' \neq p} \frac{n_p n_{p'}}{n^2} \mu_{pt}\mu_{p't} + \sum_p \frac{n_p^2}{n^2}(\frac{n_p - p_t n_p}{n_p p_t n_p} S^2_{pt} + \mu^2_{pt}) \\
&= (\sum_p \frac{n_p}{n}\mu_{pt})^2 + \sum_p \frac{n_p - p_t n_p}{n^2 p_t} S^2_{pt} \\
&= \mu_t^2 + \sum_p \frac{n_p - p_t n_p}{n^2 p_t} S^2_{pt}, \qquad (15)
\end{aligned}
$$

where I use the independence of $\hat{\mu}_{EXaM,pt}$ across p for the second equality, the fact that $E(\hat{\mu}_{EXaM,pt}|p^*(\epsilon)) = \mu_{pt}$ for the third and the fourth equalities and the first part of Lemma 6 for the fourth equality. Let $\hat{S}^2_{pt} = \frac{1}{p_t n_p - 1}\sum_{i:p_i^*(\epsilon)=p}(Y_i - \hat{\mu}_{EXaM,pt})^2 D_{it}$. By the second part of Lemma 6, $\hat{S}^2_{pt}$ is an unbiased estimator for S^2_{pt} under EXaM. Combining this with equation (15), I obtain an unbiased estimator for μ_t^2: $\hat{\theta}_{2t} = \hat{\mu}^2_{EXaM,t} - \sum_p \frac{n_p - p_t n_p}{n^2 p_t}\hat{S}^2_{pt}$. Then, by equation (14), $\hat{\theta}_{EXaM,t} = \frac{n - c_t}{(n-1)c_t}\hat{\theta}_{1t} + \frac{n(c_t - 1)}{(n-1)c_t}\hat{\theta}_{2t}$ is an unbiased estimator for $E(\hat{\mu}^2_{RCT,t}|p^{RCT})$. □

Let D_i be the set of all feasible deterministic treatment assignments for subject i, i.e.,

$$D_i \equiv \{d_i \equiv (d_{it})_t \in \{0,1\}^{m+1} | \Sigma_t d_{it} = 1\}.$$

Let $D_i^{EXaM(\epsilon)}$ and D_i^{RCT} be the sets of deterministic treatment assignments that happen with a positive probability under EXaM and RCT, respectively. That is, $D_i^{EXaM(\epsilon)} \equiv \{d_i \in D_i | \Pr(d_i|p^*(\epsilon)) > 0\}$ and $D_i^{RCT} \equiv \{d_i \in D_i | \Pr(d_i|p^{RCT}) > 0\}$, where $\Pr(d_i|p)$ is the probability that d_i occurs under experimental design p. For every $\epsilon \in [0, \bar{\epsilon})$, I have

$$D_i^{RCT} = D_i^{EXaM(\epsilon)} = D_i. \qquad (16)$$

This is because with Definition 2, for every t, with a positive probability, $D_{it} = 1$ holds both under EXaM and RCT.

With the support equivalence property (16), I am ready to show the proposition. Suppose that parameter θ is unbiasedly estimable with RCT p^{RCT} and a simple estimator $\hat{\theta}^{RCT}(Y,D) = \Sigma_i f(Y_i, D_i) + \Sigma_t g_t \hat{\mu}^2_{RCT,t}$:

$$E(\hat{\theta}^{RCT}(Y,D)|p^{RCT}) = \theta. \tag{17}$$

Consider another estimator $\hat{\theta}^{EXaM(\epsilon)}(Y,D) \equiv \Sigma_i \frac{\Pr(D_i|p^{RCT})}{\Pr(D_i|p^*(\epsilon))} f(Y_i, D_i) + \Sigma_t g_t \hat{\theta}_{EXaM,t}$. With the knowledge of the original estimator $\hat{\theta}^{RCT}(Y,D)$, it is possible to compute $\hat{\theta}^{EXaM(\epsilon)}(Y,D)$ since $\Pr(D_i|p^{RCT})$ and $\Pr(D_i|p^*(\epsilon))$ are known to the experimenter. I have:

$$\begin{aligned}
&E(\hat{\theta}^{EXaM}(Y,D)|p^*(\epsilon)) \\
&= E(\Sigma_i \frac{\Pr(D_i|p^{RCT})}{\Pr(D_i|p^*(\epsilon))} f(Y_i, D_i) + \Sigma_t g_t \hat{\theta}_{EXaM,t}|p^*(\epsilon)) \\
&= \Sigma_i \Sigma_{d_i \in D_i^{EXaM(\epsilon)}} \Pr(d_i|p^*(\epsilon)) \frac{\Pr(d_i|p^{RCT})}{\Pr(d_i|p^*(\epsilon))} f(Y_i(d_i), d_i) + \Sigma_t g_t E(\hat{\mu}^2_{RCT,t}|p^{RCT}) \\
&= \Sigma_i \Sigma_{d_i \in D_i^{RCT}} \Pr(d_i|p^{RCT}) f(Y_i(d_i), d_i) + \Sigma_t g_t E(\hat{\mu}^2_{RCT,t}|p^{RCT}) \\
&= E(\hat{\theta}^{RCT}(Y,D)|p^{RCT}) \\
&= \theta,
\end{aligned}$$

where $Y_i(d_i)$ is the value of observed outcome Y_i when $D_i = d_i$, the second equality is by Lemma 7, the third equality is by the support equivalence property (16), and the last equality is by the unbiasedness assumption (17). This means that $\hat{\theta}^{EXaM}(Y,D)$ is an unbiased estimator for θ under EXaM $p^*(\epsilon)$. The following lemma therefore completes the proof.

Lemma 8. *$\hat{\theta}^{EXaM}(Y,D)$ is a simple estimator under EXaM $p^*(\epsilon)$.*

Proof of Lemma 8. Since $\hat{\mu}_{EXaM,t} = \sum_p \frac{n_p}{n} \hat{\mu}_{EXaM,pt}$ and

$$\begin{aligned}
&\hat{S}^2_{pt} \\
&= \frac{1}{p_t n_p - 1} (\textstyle\sum_{i:p_i^*(\epsilon)=p} Y_i^2 D_{it} - 2 \sum_{i:p_i^*(\epsilon)=p} Y_i D_{it} \hat{\mu}_{EXaM,pt} + \sum_{i:p_i^*(\epsilon)=p} \hat{\mu}^2_{EXaM,pt} D_{it}) \\
&= \frac{1}{p_t n_p - 1} (\textstyle\sum_{i:p_i^*(\epsilon)=p} Y_i^2 D_{it} - 2 p_t n_p \hat{\mu}^2_{EXaM,pt} + p_t n_p \hat{\mu}^2_{EXaM,pt}) \\
&= \frac{1}{p_t n_p - 1} (\textstyle\sum_{i:p_i^*(\epsilon)=p} Y_i^2 D_{it} - p_t n_p \hat{\mu}^2_{EXaM,pt}),
\end{aligned}$$

I have

$$\hat{\theta}_{EXaM,t}$$

$$
\begin{aligned}
&= \frac{n-c_t}{(n-1)c_t}\hat{\theta}_{1t} + \frac{n(c_t-1)}{(n-1)c_t}\hat{\theta}_{2t} \\
&= \frac{n-c_t}{(n-1)c_t}\frac{1}{n}\textstyle\sum_p \frac{1}{p_t}\sum_{i:p_i^*(\epsilon)=p} Y_i^2 D_{it} + \frac{n(c_t-1)}{(n-1)c_t}(\hat{\mu}^2_{EXaM,t} - \sum_p \frac{n_p - p_t n_p}{n^2 p_t}\hat{S}^2_{pt}) \\
&= \textstyle\sum_p \sum_{i:p_i^*(\epsilon)=p} \frac{n-c_t}{(n-1)c_t n p_t} Y_i^2 D_{it} \\
&\quad + \frac{n(c_t-1)}{(n-1)c_t}\Big\{(\textstyle\sum_p \frac{n_p}{n}\hat{\mu}_{EXaM,pt})^2 - \sum_p \frac{n_p - p_t n_p}{n^2 p_t}\frac{1}{p_t n_p - 1}(\sum_{i:p_i^*(\epsilon)=p} Y_i^2 D_{it} - p_t n_p \hat{\mu}^2_{EXaM,pt})\Big\} \\
&= \textstyle\sum_p \sum_{i:p_i^*(\epsilon)=p} (\frac{n-c_t}{(n-1)c_t n p_t} - \frac{(c_t-1)(n_p - p_t n_p)}{(n-1)c_t n p_t (p_t n_p - 1)}) Y_i^2 D_{it} \\
&\quad + \frac{n(c_t-1)}{(n-1)c_t}(\textstyle\sum_p \sum_{p'} \frac{n_p n_{p'}}{n^2}\hat{\mu}_{EXaM,pt}\hat{\mu}_{EXaM,p't} + \sum_p \frac{(n_p - p_t n_p)n_p}{n^2(p_t n_p - 1)}\hat{\mu}^2_{EXaM,pt}) \\
&= \textstyle\sum_p \sum_{i:p_i^*(\epsilon)=p} a_{1pt} Y_i^2 D_{it} + \sum_p \sum_{p'\neq p} \frac{(c_t-1)n_p n_{p'}}{(n-1)c_t n}\hat{\mu}_{EXaM,pt}\hat{\mu}_{EXaM,p't} \\
&\quad + \sum_p \frac{n(c_t-1)}{(n-1)c_t}(\frac{n_p^2}{n^2} + \frac{(n_p - p_t n_p)n_p}{n^2(p_t n_p - 1)})\hat{\mu}^2_{EXaM,pt} \\
&= \textstyle\sum_i a_{1p_i^*(\epsilon)t} Y_i^2 D_{it} + \sum_p \sum_{p'} a_{2pp't}\hat{\mu}_{EXaM,pt}\hat{\mu}_{EXaM,p't},
\end{aligned}
$$

where

$$
\begin{aligned}
a_{1pt} &= \frac{(n-1)p_t n_p - (c_t-1)n_p - n + c_t}{(n-1)c_t n p_t (p_t n_p - 1)} \\
a_{2pp't} &= \begin{cases} \dfrac{(c_t-1)n_p n_{p'}}{(n-1)c_t n} & \text{if } p \neq p' \\ \dfrac{(c_t-1)n_p^2(p_t n_p - p_t)}{(n-1)c_t n(p_t n_p - 1)} & \text{if } p = p'. \end{cases}
\end{aligned}
$$

It follows that

$$
\begin{aligned}
&\hat{\theta}^{EXaM}(Y,D) \\
&= \textstyle\sum_i \frac{\Pr(D_i|p^{RCT})}{\Pr(D_i|p^*(\epsilon))} f(Y_i, D_i) + \sum_t g_t \hat{\theta}_{EXaM,t} \\
&= \textstyle\sum_i [\frac{\Pr(D_i|p^{RCT})}{\Pr(D_i|p^*(\epsilon))} f(Y_i, D_i) + \sum_t g_t a_{1p_i^*(\epsilon)t} Y_i^2 D_{it}] + \sum_t g_t \sum_p \sum_{p'} a_{2pp't}\hat{\mu}_{EXaM,pt}\hat{\mu}_{EXaM,p't} \\
&= \textstyle\sum_i f^*(Y_i, D_i) + \sum_t \sum_p \sum_{p'} g_{tpp'}\hat{\mu}_{EXaM,pt}\hat{\mu}_{EXaM,p't},
\end{aligned}
$$

where $f^*(Y_i, D_i) = \dfrac{\Pr(D_i|p^{RCT})}{\Pr(D_i|p^*(\epsilon))} f(Y_i, D_i) + \sum_t g_t a_{1p_i^*(\epsilon)t} Y_i^2 D_{it}$ and $g_{tpp'} = g_t a_{2pp't}$. Therefore,

$\hat{\theta}^{EXaM}(Y, D)$ is a simple estimator under EXaM $p^*(\epsilon)$. □

Proof of Proposition 5

To show the mean part, recall I define $N_{pt} \equiv \sum_i 1\{p_i^*(\epsilon) = p\}D_{it}$ as a random variable that stands for the number of subjects with propensity vector p and assigned to treatment t. Denote the realization of N_{pt} by $n_{pt} \equiv \sum_i 1\{p_i^*(\epsilon) = p\}d_{it}$. By Lemma 10, every feasible treatment assignment occurs equally likely conditional on (N_{pt}) so that that for every p, t and i with $p_i^*(\epsilon) = p$,

$$E(D_{it}|(N_{pt}) = (n_{pt}), p^*(\epsilon) = p) = \frac{n_{pt}}{n_p}. \tag{18}$$

I therefore have

$$\begin{aligned}
&E(\hat{\beta}_t|(N_{pt}) = (n_{pt}), p^*(\epsilon) = p) \\
&= E(\Sigma_p \delta_p \hat{\beta}_{pt}|(N_{pt}) = (n_{pt}), p^*(\epsilon) = p) \\
&= \Sigma_p \delta_p E(\hat{\beta}_{pt}|(N_{pt}) = (n_{pt}), p^*(\epsilon) = p) \\
&= \Sigma_p \delta_p E\big[\Sigma_i 1\{p_i^*(\epsilon) = p\}\big(\frac{D_{it}Y_i(t)}{N_{pt}} - \frac{D_{it_0}Y_i(t_0)}{N_{pt_0}}\big)|(N_{pt}) = (n_{pt}), p^*(\epsilon) = p\big] \\
&= \Sigma_p \delta_p \Sigma_i 1\{p_i^*(\epsilon) = p\}\big(\frac{E(D_{it}|(N_{pt}) = (n_{pt}), p^*(\epsilon) = p)Y_i(t)}{n_{pt}} - \frac{E(D_{it_0}|(N_{pt}) = (n_{pt}), p^*(\epsilon) = p)Y_i(t_0)}{n_{pt_0}}\big) \\
&= \Sigma_p \delta_p \Sigma_i 1\{p_i^*(\epsilon) = p\}\big(\frac{(n_{pt}/n_p)Y_i(t)}{n_{pt}} - \frac{(n_{pt_0}/n_p)Y_i(t_0)}{n_{pt_0}}\big) \\
&= \Sigma_p \delta_p \frac{1}{n_p} \Sigma_i 1\{p_i^*(\epsilon) = p\}(Y_i(t) - Y_i(t_0)) \\
&= \Sigma_p \delta_p CATE_{pt},
\end{aligned}$$

where I use equation (18) for the fifth equality. By the law of iterated expectations, I conclude

$$\begin{aligned}
&E(\hat{\beta}_t|p^*(\epsilon)) \\
&= E[E(\hat{\beta}_t|(N_{pt}) = (n_{pt}), p^*(\epsilon) = p)|p^*(\epsilon) = p] \\
&= E[\Sigma_p \delta_p CATE_{pt}|p^*(\epsilon) = p] \\
&= \Sigma_p \delta_p CATE_{pt}.
\end{aligned}$$

For the variance part, I prove the general version given in Appendix A.1.1. Define $\mathcal{N}$ as the set of all (n_{pt}) that satisfy the following:

- $n_{pt} = \underline{n}_{pt}$ for all p and t such that $p_t n_p \in \mathbb{N}$.
- $n_{pt} \in \{\underline{n}_{pt}, \underline{n}_{pt} + 1\}$ for all p and t such that $p_t n_p \notin \mathbb{N}$.

- $\sum_t n_{pt} = n_p$ for all p.
- $\sum_p n_{pt} = c_t$ for all t.

I also define $D(n_{pt})$ as the set of deterministic treatment assignments where the realization of (N_{pt}) is (n_{pt}):

$$D(n_{pt}) \equiv \{d \in \{0,1\}^{n\times m} | \sum_t d_{it} = 1 \text{ for every } i \text{ and } \sum_i 1\{p_i^*(\epsilon) = p\} d_{it} = n_{pt} \text{ for every } p \text{ and } t\}.$$

The way of drawing deterministic treatment assignments in Definition 2 in Appendix A.1.1 satisfies the following properties.

Lemma 9 (Small Support)**.** *The support of* (N_{pt}) *is included by* $\mathcal{N}$*.*

Lemma 10 (Conditional Uniformity)**.** *For all* (n_{pt}) *in the support of* (N_{pt})*,*

$$Pr(D = d | (N_{pt}) = (n_{pt}), p^*(\epsilon)) = \begin{cases} |D(n_{pt})|^{-1} & \text{if } d \in D(n_{pt}) \\ 0 & \text{otherwise,} \end{cases}$$

where $|D(n_{pt})| = \prod_p \prod_{j=0}^{m-1} \binom{\sum_{j'=0}^{j+1} n_{pt_{j'}}}{\sum_{j'=0}^{j} n_{pt_{j'}}}$.

For notational simplicity, I make conditioning on $p^*(\epsilon)$ implicit. By the law of total variance, $V(\hat{\beta}_t)$ can be written as:

$$V(\hat{\beta}_t) = E(V(\hat{\beta}_t | (N_{pt}))) + V(E(\hat{\beta}_t | (N_{pt}))).$$

As I show above, $E(\hat{\beta}_t | (N_{pt})) = \sum_p \delta_p CATE_{pt}$, implying $V(E(\hat{\beta}_t | (N_{pt}))) = 0$. Thus

$$V(\hat{\beta}_t) = E(V(\hat{\beta}_t | (N_{pt}))). \tag{19}$$

To show that $E(V(\hat{\beta}_t | (N_{pt})))$ is equal to the right-hand side of equation (12), I introduce a lemma.

Lemma 11. *Under Lemma 10, for all* (n_{pt}) *in the support of* (N_{pt})*,*

$$V(\hat{\beta}_t | (N_{pt}) = (n_{pt})) = \sum_p \delta_p^2 \Big(\frac{S_{pt}^2}{n_{pt}} + \frac{S_{pt_0}^2}{n_{pt_0}} - \frac{S_{ptt_0}^2}{n_p} \Big).$$

Proof of Lemma 11. By Lemma 10, treatment assignments are independent across subpopulations with different propensities conditional on (N_{pt}). Then, $\hat{\beta}_{pt}$ is independent across p

conditional on (N_{pt}). Hence,

$$\begin{aligned} V(\hat{\beta}_t|(N_{pt}) = (n_{pt})) &= V(\sum_p \delta_p \hat{\beta}_{pt}|(N_{pt}) = (n_{pt})) \\ &= \sum_p \delta_p^2 V(\hat{\beta}_{pt}|(N_{pt}) = (n_{pt})). \end{aligned}$$

It is therefore enough to show that $V(\hat{\beta}_{pt}|(N_{pt}) = (n_{pt})) = \frac{S_{pt}^2}{n_{pt}} + \frac{S_{pt0}^2}{n_{pt0}} - \frac{S_{ptt0}^2}{n_p}$. For notational simplicity, I make conditioning on $(N_{pt}) = (n_{pt})$ implicit. Let I^{ptt_0} be a random set of subjects with propensity vector p and assigned to either treatment t or t_0, i.e.,

$$I^{ptt_0} \equiv \{i|p_i^*(\epsilon) = p \text{ and } D_{it} + D_{it_0} = 1\}.$$

I^{ptt_0} takes on $\binom{n_p}{n_{pt} + n_{pt0}}$ values equally likely by Lemma 10. By the law of total variance, $V(\hat{\beta}_{pt})$ can be written as:

$$V(\hat{\beta}_{pt}) = E(V(\hat{\beta}_{pt}|I^{ptt_0})) + V(E(\hat{\beta}_{pt}|I^{ptt_0})). \tag{20}$$

Conditional on $I^{ptt_0} = I$, the randomness in $\hat{\beta}_{pt}$ comes from the randomness in choosing n_{pt} subjects assigned to treatment t and n_{pt_0} subjects assigned to treatment t_0 from the set I of $n_{pt} + n_{pt_0}$ subjects. Every combination occurs with equal probability, so the standard results of binary-treatment RCT (Theorems 6.1 and 6.2 in Imbens and Rubin (2015)) apply:

$$\begin{aligned} E(\hat{\beta}_{pt}|I^{ptt_0} = I) &= \frac{1}{n_{pt} + n_{pt_0}} \sum_{i \in I} (Y_i(t) - Y_i(t_0)), \\ V(\hat{\beta}_{pt}|I^{ptt_0} = I) &= \frac{S_{pt|I}^2}{n_{pt}} + \frac{S_{pt_0|I}^2}{n_{pt_0}} - \frac{S_{ptt_0|I}^2}{n_{pt} + n_{pt_0}}, \end{aligned}$$

where $S_{pt|I}^2$, $S_{pt0|I}^2$ and $S_{ptt0|I}^2$ are the variances of $Y_i(t)$, $Y_i(t_0)$ and $Y_i(t) - Y_i(t_0)$, respectively, in the set of subjects I. Regarding $n_p, n_{pt} + n_{pt_0}$, and $Y_i(t) - Y_i(t_0)$ as performing the roles of n, m, and X_i, respectively, I use Lemma 6 to get

$$\begin{aligned} & V(E(\hat{\beta}_{pt}|I^{ptt_0} = I)) \\ &= V(\frac{1}{n_{pt} + n_{pt_0}} \textstyle\sum_{i \in I} (Y_i(t) - Y_i(t_0))) \\ &= \frac{n_p - n_{pt} - n_{pt_0}}{n_p(n_{pt} + n_{pt_0})} S_{ptt_0}^2, \end{aligned}$$

$$
\begin{aligned}
&E(V(\hat{\beta}_{pt}|I^{ptt_0}=I)) \\
&= \frac{E(S^2_{pt|I})}{n_{pt}} + \frac{E(S^2_{pt_0|I})}{n_{pt_0}} - \frac{E(S^2_{ptt_0|I})}{n_{pt}+n_{pt_0}} \\
&= \frac{S^2_{pt}}{n_{pt}} + \frac{S^2_{pt_0}}{n_{pt_0}} - \frac{S^2_{ptt_0}}{n_{pt}+n_{pt_0}}.
\end{aligned}
$$

Combining these with equation (20), we have $V(\hat{\beta}_{pt}) = \frac{S^2_{pt}}{n_{pt}} + \frac{S^2_{pt_0}}{n_{pt_0}} - \frac{S^2_{ptt_0}}{n_p}$. □

By Lemma 9, N_{pt} can take on either $\underline{n}_{pt}$ or $\underline{n}_{pt}+1$. Since N_{pt} has expectation $p_t n_p$, the marginal distribution for each N_{pt} must be

$$
Pr(N_{pt}=n_{pt}) = \begin{cases} 1-p_t n_p + \underline{n}_{pt} & \text{if } n_{pt}=\underline{n}_{pt} \\ p_t n_p - \underline{n}_{pt} & \text{if } n_{pt}=\underline{n}_{pt}+1 \\ 0 & \text{otherwise.} \end{cases} \tag{21}
$$

Using equation (19), Lemma 11, and equation (21), we have

$$
\begin{aligned}
&V(\hat{\beta}_t) \\
&= E\Big\{\textstyle\sum_p \delta_p^2 \Big(\frac{S^2_{pt}}{N_{pt}} + \frac{S^2_{pt_0}}{N_{pt_0}} - \frac{S^2_{ptt_0}}{n_p}\Big)\Big\} \\
&= \textstyle\sum_p \delta_p^2 \Big\{\sum_{t'\in\{t_0,t\}} E\Big(\frac{S^2_{pt'}}{N_{pt'}}\Big) - \frac{S^2_{ptt_0}}{n_p}\Big\} \\
&= \textstyle\sum_p \delta_p^2 \Big\{\sum_{t'\in\{t_0,t\}} \Big[\Big(\frac{S^2_{pt'}}{\underline{n}_{pt'}}\Big)(1-p_{t'}n_p+\underline{n}_{pt'}) + \Big(\frac{S^2_{pt'}}{\underline{n}_{pt'}+1}\Big)(p_{t'}n_p-\underline{n}_{pt'})\Big] - \frac{S^2_{ptt_0}}{n_p}\Big\}.
\end{aligned}
$$

Proof of Equation (7)

I prove equation (7) with two lemmas below.

Lemma 12. *$E(\hat{B}_t|p^*(\epsilon)) = \Sigma_p \lambda_p CATE_{pt}$ for all t where $\hat{B}_t$ is the OLS estimate of B_t in this regression:*

$$
Y_i = \Sigma_{t=t_1}^{t_m} B_t D_{it} + \Sigma_p C_p 1\{(p_i^*(\epsilon)=p\} + E_i. \tag{22}
$$

Proof of Lemma 12. I reparametrize the regression as follows with (B_t, D_p), where $D_p \equiv C_p + \Sigma_{t=t_1}^{t_m} B_t p_t$.

$$
Y_i = \Sigma_{t=t_1}^{t_m} B_t (D_{it} - p_{it}^*(\epsilon)) + \Sigma_p D_p 1\{(p_i^*(\epsilon)=p\} + E_i. \tag{23}
$$

This reparametrization does not change $\hat{B}_t$. Note also that Y_i can be written as follows.

$$Y_i = \Sigma_p 1\{(p_i^*(\epsilon) = p\}Y_p(0) + \Sigma_p \Sigma_{t=t_1}^{t_m} 1\{(p_i^*(\epsilon) = p\}CATE_{pt}D_{it} + \mu_i, \tag{24}$$

where $Y_p(0) \equiv \dfrac{\Sigma_i 1\{(p_i^*(\epsilon) = p\}Y_i(0)}{\Sigma_i 1\{(p_i^*(\epsilon) = p\}}$, and where $\Sigma_i 1\{(p_i^*(\epsilon) = p\}\mu_i = 0$ for every p. Therefore, the OLS estimates $(\hat{B}_t, \hat{D}_p)$ of (B_t, D_p) in regression (23) can be written as follows.

$$\begin{aligned}
(\hat{B}_t, \hat{D}_p) &= \underset{(B_t, D_p)}{\arg\min}\, \Sigma_i[\Sigma_p 1\{(p_i^*(\epsilon) = p\}Y_p(0) + \Sigma_p \Sigma_{t=t_1}^{t_m} 1\{(p_i^*(\epsilon) = p\}CATE_{pt}D_{it} \\
&\quad - \Sigma_{t=t_1}^{t_m} B_t(D_{it} - p_{it}^*(\epsilon)) - \Sigma_p D_p 1\{(p_i^*(\epsilon) = p\}]^2 \\
&= \underset{(B_t, D_p)}{\arg\min}\, \Sigma_i[\Sigma_p 1\{(p_i^*(\epsilon) = p\}(Y_p(0) - D_p + \Sigma_{t=t_1}^{t_m} CATE_{pt}D_{it}) - \Sigma_{t=t_1}^{t_m} B_t(D_{it} - p_{it}^*(\epsilon)]^2 \\
&= \underset{(B_t, D_p)}{\arg\min}\, \Sigma_i[\{\Sigma_p 1\{(p_i^*(\epsilon) = p\}(Y_p(0) - D_p + \Sigma_{t=t_1}^{t_m} CATE_{pt}D_{it}\}^2 \\
&\quad - 2\Sigma_p 1\{(p_i^*(\epsilon) = p\}(Y_p(0) - D_p + \Sigma_{t=t_1}^{t_m} CATE_{pt}D_{it})\Sigma_{t=t_1}^{t_m} B_t(D_{it} - p_{it}^*(\epsilon)) \\
&\quad + \{\Sigma_{t=t_1}^{t_m} B_t(D_{it} - p_{it}^*(\epsilon)\}^2] \\
&= \underset{(B_t, D_p)}{\arg\min}\, \Sigma_i[\{\Sigma_p 1\{(p_i^*(\epsilon) = p\}(Y_p(0) - D_p + \Sigma_{t=t_1}^{t_m} CATE_{pt}D_{it})\}^2 \\
&\quad - 2\Sigma_p 1\{(p_i^*(\epsilon) = p\}\Sigma_{t=t_1}^{t_m} CATE_{pt}D_{it}\Sigma_{t=t_1}^{t_m} B_t(D_{it} - p_{it}^*(\epsilon)) + \{\Sigma_{t=t_1}^{t_m} B_t(D_{it} - p_{it}^*(\epsilon))\}^2]
\end{aligned}$$

because $\Sigma_i(D_{it} - p_{it}^*(\epsilon)) = 0$. Minimizing this over B_t leads to

$$\hat{B}_t = \frac{\Sigma_i \Sigma_p 1\{(p_i^*(\epsilon) = p\}CATE_{pt}D_{it}(D_{it} - p_{it}^*(\epsilon))}{\Sigma_i(D_{it} - p_{it}^*(\epsilon))^2}$$

Because $P(D_{it} = 1) = \dfrac{\Sigma_p \Sigma_i 1\{(p_i^*(\epsilon) = p\}p_{it}^*(\epsilon)}{n}$ and

$$P(p_i^*(\epsilon) = p | D_{it} = 1) = \frac{\Sigma_i 1\{(p_i^*(\epsilon) = p\}p_{it}^*(\epsilon)}{\Sigma_q \Sigma_i 1\{(p_i^*(\epsilon) = q\}p_{it}^*(\epsilon)},$$

it follows that the numerator is equal to $\Sigma_p p_t(1 - p_t)\delta_p CATE_{pt}$ and that the denominator is equal to $\Sigma_p p_t(1 - p_t)\delta_p$. This implies that $E(\hat{B}_t | p^*(\epsilon)) = \Sigma_p \lambda_p CATE_{pt}$. □

Lemma 13. *$\hat{B}_t = \hat{b}_t$ for any t and any realization of treatment assignment D_{it}.*

Proof of Lemma 13. The OLS estimates of (22) can be obtained by regressing each of Y_i and D_{it} on the propensity score controls and then using the residuals from these regressions as

the dependent and independent variables for a bivariate regression that omits the propensity score controls. Consider the auxiliary regressions that produce these residualized variables: they have D_{it} on the left hand side, with a saturated model for $p_i^*(\epsilon)$ on the right. By the law of iterated expectations, the conditional expectation function associated with this auxiliary regression is therefore

$$E[D_{it}|p_i^*(\epsilon)] = p_{it}^*(\epsilon).$$

In other words, the conditional expectation function depends only on $p_{it}^*(\epsilon)$. Moreover, because I use a saturated model for the own-score $p_i^*(\epsilon)$, the conditional expectation function $E[D_{it}|p_i^*(\epsilon)]$ is linear in regressors, so it and the associated auxiliary regression function coincide. Therefore, regression (6), which additively separably and linearly controls for $p_{it}^*(\epsilon)$'s, produces the same estimate as regression (22). □

Proof of Proposition 6

By Proposition 2, there is no other experimental design (p_{it}) with $p_{it} \in [\epsilon, 1-\epsilon]$ for all subject i and treatment t and such that $\Sigma_t p_{it} w'_{it} \geq \Sigma_t p_{it}^{*o}(\epsilon) w'_{it}$ for all i and $\Sigma_t p_{it} e'_{ti} \geq \Sigma_t p_{it}^{*o}(\epsilon) e'_{ti}$ for all i with at least one strict inequality. w'_{it} and e'_{ti} are consistent with ordinal $\succsim_i$ and $\succsim_t$, respectively. Therefore, there is no other experimental design (p_{it}) such that for all cardinal WTP w_{it} consistent with ordinal $\succsim_i$ and all cardinal predicted effects e_{ti} consistent with ordinal $\succsim_t$, I have $\Sigma_t p_{it} w_{it} \geq \Sigma_t p_{it}^{*o}(\epsilon) w_{it}$ for all i and $\Sigma_t p_{it} e_{ti} \geq \Sigma_t p_{it}^{*o}(\epsilon) e_{ti}$ for all i with at least one strict inequality.

A.2 Empirical Appendix

A.2.1 Why Subject Welfare: Data

Table 1 and Appendix Tables A.1-A.3 are based on data I assemble from the WHO International Clinical Trials Registry Platform (ICTRP) at `http://www.who.int/ictrp/en/`, retrieved in October 2017. I first use the "date of registration" to define the year associated with each trial. Starting from the universe of trials registered between January 1st 2007 to May 31st 2017, I exclude outlier trials with registered sample size greater than 5 millions. Some trials come with sample size classified as "Not Specified." I set their sample size as zero. This makes my total sample size calculation conservative. For a trial that does not have a well-defined trial phase, I classify its trial phase as "Not Specified." Finally, for each trial, I define its "Geographical Region" according to which country runs the registry including that trial. Many registries like ClinicalTrial.gov recruit subjects in multiple countries under the same trial ID, making it challenging to pin down the physical location of each trial.

Appendix Tables A.4, A.5, A.6 are based on data I assemble from the American Economic Association's registry (AEA registry) for randomized controlled trials at `https://www.socialscienceregistry.org`, retrieved on May 27th, 2017. From the AEA registry, I obtain information about each experiment such as the sample size, the year when the experiment is conducted, the country where the experiment is conducted, registered keywords, and the randomization unit. When some information is missing, I manually enter it by referring to accompanying documents such as experimental design descriptions and abstracts. I classify an item as "Not specified" when I cannot specify it even after the manual procedure. When the sample size of an experiment is unspecified, I set the sample size as zero. This makes my total sample size calculation conservative. I use the "starting date of experiment" to define the year associated with each trial. Finally, for each trial, I define its "Geographical Region" according to in which country the experiment is conducted. I include all registered experiments conducted during 2007-2017 period.

A.2.2 Empirical Application: Data

For the OLS regressions in Table 3, I impose the same sample restriction as Kremer et al. and exclude the following children: children not at Intent-to-Treat springs, i.e., springs found to be nonviable after treatment random assignment, children in households that receive water guards in 2007, children not in representative households (defined as households that are named at least twice by all users of a given spring, every time survey enumerators ask spring users at a spring to name households that also use the same spring and when enumerators ask three or four households located nearest to a spring to name spring users), children above age 3 at baseline and children above age 3 when they join the sample in later rounds, children whose anthropometric (weight, height, BMI) and age data are flagged as having serious error, and children in households with missing data on whether they use the identified spring exclusively or use multiple springs.

A.2.3 Treatment Effects and Preferences: Details

Estimation of Mixed Logit WTP Model (Table 4)

With the random utility function (11), choice likelihoods take the following form (Train (2003), chapter 6):

$$P(o_{ijt} = 1|\theta, \gamma_1, \delta_j) = \int_{(\beta_i, c_i)} \frac{\exp\left(\beta_i + \gamma_1 X_i)T_{jt} - c_i D_{ij} + \delta_j\right.}{\sum_{h \in H} \exp\left(\beta_i + \gamma_1 X_i)T_{ht} - c_i D_{ih} + \delta_h\right.} f(\beta_i, c_i|\theta) d(\beta_i, c_i) \quad (25)$$

where $o_{ijt} \in \{0, 1\}$ is the indicator that household i chooses source j among alternatives $h \in H$ in trip t; $f(\beta_i, c_i|\theta)$ is the mixing distribution parametrized by θ. $f(\beta_i, c_i|\theta)$ is taken to be the normal distribution for the spring protection treatment coefficient β_i and the triangular distribution (restricted to be nonnegative) for the distance coefficient c_i. I maximize a simulation approximation of the joint likelihood $\Sigma_{ijt} P(o_{ijt} = 1|X)$ with respect to θ, γ_1, and δ_j, producing maximum simulated likelihood estimates $\hat{\theta}, \hat{\gamma}_1$, and $\hat{\delta}_j$.

Simulation of Heterogeneous WTP (Figure 1b)

I create Figure 1b with parametric bootstrap below.

(1) Simulate $\mu^T \sim N(\widehat{\mu^T}, SE(\widehat{\mu^T}))$ and $\sigma^T \sim N(\widehat{\sigma^T}, SE(\widehat{\sigma^T}))$ for treatment coefficient parameters across household groups.

(2) Simulate $\theta^D \sim Triangular(\widehat{\theta^D})$ for distance coefficient across household group.[39]

(3) For each simulated value of μ^T and σ^T, draw the treatment coefficient for each household group from $N(\mu^T, \sigma^T)$. Call this θ^T. The θ^D from step 2 is the distance coefficient because the distribution relies on only one parameter. Find the ratio of θ^T to θ^D. Multiplying the figure by -1/0.38 generates the ratio of treatment to distance coefficients, where -1 is the non-negative correction multiplier and 0.38 is the correlation across survey rounds in the reported walking distance to the reference spring and is taken to be the size of measurement error from recall error. I do this coefficient inflation following Kremer et al.

(4) I multiply the ratio of step 3 by $32 \times 52/(60 \times 8)$ in order to get the WTP measure, with total number of working days taken to walk to the spring in a year as the unit.

A.2.4 EXaM vs RCT: Details

Implementing EXaM

EXaM assigns subject i to treatment t with probability $p^*_{it}(\epsilon)$, which I define as probabilities obeying the equilibrium conditions in Definition 2. To implement random assignment with

[39]The mixed logit result of the distance coefficient is constrained to be nonnegative, because the utility function assumes u_{ijt} to be a function of $-C_i D_{ij}$ with the minus sign already reflected. But the distribution from which the distance is drawn is not constrained in terms of values. It is only constrained in terms of the mean and the spread being the same: $a + a * t$ where t is between -1 and 1, and a is the mean and the spread. The results of these logit regressions could therefore be negative. (Kremer et al. also reports negative distance coefficients for all regressions in Table VI.) Presumably Kremer et al. then took the liberty to constrain these negative distance coefficients to be multiplied by -1 in their code in order to make them nonnegative, which results in positive valuation of workdays and dollar time values.

$p^*_{it}(\epsilon)$, I need to compute $p^*_{it}(\epsilon)$ with a constructive algorithm. In this section, I describe the details of the algorithm I use for the empirical execution of EXaM in Section 7.3. I first define subroutines and then call them together at the end to perform the main computation.

Algorithm 1 Experimental Design as Market Design (EXaM)

Input: n the number of subjects, m the number of treatments, $(c_t)_t \in \mathbb{N}$ treatment t's capacity with $\sum_t c_t = n$, $(w_{it})_{i,t}$ subject i's WTP for treatment t, $(e_{ti})_{i,t}$ treatment t's predicted treatment effect for subject i, b the budget constraint, $0 < \epsilon$ the lower bound on treatment probabilities.

Output: $(p^*_{it})_{i,t}$ treatment t's assignment probability for subject i, $(\alpha^*_t, \beta^*_t)_t$ parameters determining treatment t's equilibrium price of the form $\pi^*_{te} = \alpha^*_t e + \beta^*_t$.

function InitialAlpha( )
  for each t **do**
    $\alpha_t \leftarrow$ generate random number $\sim$ Uniform$(-b, 0)$ ▷ set the initial value of α_t
  return $(\alpha_t)_t$ ▷ return an mt-dimensional vector
function InitBeta( )
  for each t **do**
    $\beta_t \leftarrow$ generate random number $\sim$ Uniform$(-b, b)$ ▷ set the initial value of β_t
  return $(\beta_t)_t$ ▷ return an m-dimensional vector

function Price$((\alpha_t)_t, (\beta_t)_t)$ ▷ get the price of treatment t
  for each i, t **do**
    $\pi_{te_{ti}} = \alpha_t e_{ti} + \beta_t$
  return $(\pi_{te_{ti}})_{it}$ ▷ return the $n \times m$ price matrix)

function Demand$(\epsilon, (\pi_{te_{ti}})_{it})$ ▷ get the subject i's demand for treatment t
  for Each i **do** ▷ perform linear programming utility maximization for each subject i
    $(p_{it})_t \leftarrow \arg\max_{(p_{it})_t} \sum_t w_{it} p_{it}$
      s.t. $\sum_t \pi_{te_{ti}} p_{it} \leq b$, $\sum_t p_{it} = 1$, $\epsilon \leq p_{it} \leq 1 - \epsilon$
  return $(p_{it})_{it}$ ▷ return the $n \times m$ demand matrix)

function ExcessDemand$((p_{it})_{it})$ ▷ get the excess demand for treatment t
  for each t **do**
    $d_t \leftarrow \sum_i p_{it} - c_t$
  return $(d_t)_t$ ▷ return the m-dimensional excess demand vector)

function ClearingError$((d_t)_t)$ ▷ get the market clearing error
  if $d_t < 0$ for all t **then**
    return 0
  else
    error $\leftarrow \sqrt{\sum_t d_t^2} / \sum_t c_t$
    return error ▷ return the market clearing error

$\delta_\alpha \leftarrow 0.75$ ▷ scaling factor for α_t's to set new prices
$\delta_\beta \leftarrow b/50$ ▷ scaling factor for β_t's to set new prices

function BetaNew($(\alpha_t)_t$, $(d_t)_t$) ▷ recalibrate β_t's to set new prices
    for each t **do**
        $\beta_t^{new} \leftarrow \beta_t + d_t\delta_\beta$
    return $(\beta_t^{new})_t$

function ClearMarket( ) ▷ the main function
    $(\alpha_t)_t \leftarrow$ InitialAlpha( )
    $(\beta_t)_t \leftarrow$ InitBeta( )
    $(\pi_{te_{ti}})_{it} \leftarrow$ Price$((\alpha_t)_t, (\beta_t)_t)$
    $(p_{it})_{it} \leftarrow$ Demand$((\pi_{it})_{it})$
    $(d_t)_t \leftarrow$ ExcessDemand$((p_{it})_{it})$
    error $\leftarrow$ ClearingError$((d_t)_t)$
    $\text{error}_{min} \leftarrow$ error ▷ initialize the min of clearing error
    ClearingThreshold $\leftarrow 0.01$ ▷ threshold for market clearing error
    IterationThreshold $\leftarrow 10$ ▷ threshold for iteration times
    iterations $\leftarrow 0$ ▷ initialize iteration time count
    while True **do**
    if iterations $>$ IterationThreshold **then**
        $(\alpha_t)_t \leftarrow$ InitialAlpha( ) ▷ start new equilibrium research
        $(\beta_t)_t \leftarrow$ InitBeta( )
        $Iterations \leftarrow 0$
    else
        $(\beta_t)_t \leftarrow$ BetaNew$((\beta_t)_t, (d_t)_t)$
    $(\pi_{te_{ti}})_{it} \leftarrow$ Price$((\alpha_t)_t, (\beta_t)_t)$
    $(p_{it})_{it} \leftarrow$ Demand$(\epsilon, (\pi_{te_{ti}})_{it})$
    $(d_t)_t \leftarrow$ ExcessDemand$((p_{it})_{it})$
    error $\leftarrow$ ClearingError$((d_t)_t)$
    if error $<$ error_{min} **then**
        $\text{error}_{min} \leftarrow$ error
        $(\alpha_t^*)_t \leftarrow (\alpha_t)_t$ ▷ the new prices reduce the error
        $(\beta_t^*)_t \leftarrow (\beta_t)_t$
        $(p_{it}^*)_{it} \leftarrow (p_{it})_{it}$
    if error_{min} $<$ ClearingThreshold **then**
        break
    iterations += 1
    return $((p_{it}^*)_{it}, (\alpha_t^*)_t, (\beta_t^*)_t, \text{error}_{min})$ ▷ return the outputs

Information (Figure 3)

After simulating treatment effects and WTP (e_{it_1}, w_{it_1}) following the procedure in Appendix A.2.3 and running the EXaM algorithm to get treatment assignment probability $p^*_{it_1}(\epsilon)$, I use $p^*_{it_1}(\epsilon)$ to draw a final deterministic treatment assignment under EXaM, i.e., $D_i \equiv 1\{i \text{ is assigned to } t_1\}$, using the following rules. Within each sampling step $s \in [1, 1540]$, randomly pick household i_s and draw $D_{i_s} = 1$ with its treatment assignment probability $p^*_{i_s t_1}(\epsilon)$. Continue until I reach treatment and control capacities. That is, if $\sum_{s'=1}^{s} D_{i_{s'}} < c_{t_1}$, then move on to step $s+1$. I stop the simulation if I exhaust any of the capacity constraints, i.e., if $\sum_{s'=1}^{s} D_{i_{s'}} = c_{t_1}$, then stop simulation and $D_j = 0$ for $j \neq i_1, i_2, i_3, ..., i_s$, or if $\sum_{s'=1}^{s} D_{i_{s'}} = c_{t_0}$, then stop simulation and $D_j = 1$ for $j \neq i_1, i_2, i_3, ..., i_s$, whichever comes first. Computationally, I achieve the above rules utilizing the random permutation algorithm:

(1) Create a random sequence of subject-picking. Specifically, I draw a number from $R_i \sim_{iid} U[0,1]$ for each household i and sort the random number in ascending order.

(2) Assign each household i to the treatment t_1 if $R_i \leq p^*_{it_1}(\epsilon)$

(3) By the random sequence of step 1, calculate the cumulative sum of treatment assignment and control assignment.

(4) If the cumulative sum of treatment assignment reaches its capacity of 663, assign the control t_0 to subjects in the remaining part of the sequence, regardless of the assignment in step 2. If the cumulative sum of control assignment reaches its capacity of 887, assign the treatment t_1 to subjects in the remaining part of the sequence, regardless of the assignment in step 2.

The treatment assignment procedure for RCT is the same except that the treatment assignment probability is $p^{RCT}_{it_1} = .56 (= 877/1540)$ and the same for everybody.

A.2.5 Additional Tables and Figures

Table A.1: Magnitude of a Part of the Clinical Trial Industry: Details 1

w/o >5 Million Sample Sizes	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2007-2017
# of Clinical Trials Registered	16,502	23,349	24,468	26,803	25,415	26,794	29,406	31,707	34,854	39,793	17,506	296,597
Sum of Sample Sizes	13,965,001	21,737,552	15,076,229	29,863,810	28,395,906	31,615,304	34,383,227	38,990,898	61,796,652	72,113,892	19,964,109	367,902,580
w/o >1 Million Sample Sizes	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2007-2017
# of Clinical Trials Registered	16,495	23,346	24,467	26,801	25,411	26,790	29,400	31,699	34,843	39,781	17,504	296,543
Sum of Sample Sizes	7,229,849	14,786,492	13,146,328	25,781,618	19,471,190	21,215,671	21,998,112	25,482,511	34,575,586	43,682,197	16,936,550	246,041,256
w/o >0.5 Million Sample Sizes	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2007-2017
# of Clinical Trials Registered	16,495	23,344	24,464	26,793	25,405	26,785	29,397	31,695	34,831	39,766	17,498	296,478
Sum of Sample Sizes	7,229,849	13,176,883	11,026,328	19,578,067	14,871,617	17,495,905	19,946,994	22,653,444	24,884,051	32,107,344	11,812,087	195,517,722
w/o >0.1 Million Sample Sizes	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2007-2017
# of Clinical Trials Registered	16,495	23,333	24,458	26,768	25,394	26,765	29,370	31,653	34,797	39,718	17,482	296,231
Sum of Sample Sizes	7,229,849	10,890,648	9,939,604	14,204,355	12,486,159	13,187,175	13,610,722	13,975,686	16,240,202	22,109,833	8,662,407	142,536,640

Notes: This extends Table 1 and details its figures with different sample definitions. This table provides summary statistics of clinical trials registered in the WHO International Clinical Trials Registry Platform (ICTRP, `http://www.who.int/ictrp/en/`, retrieved in October 2017). The sample consists of clinical trials registered there between January 1st, 2007 to May 30th, 2017. See Section 2 for discussions about this exhibit and Appendix A.2.1 for the computational procedure.

Table A.2: Magnitude of a Part of the Clinical Trial Industry: Details 2

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2007-2017
# of Clinical Trials Registered	16,502	23,349	24,468	26,803	25,415	26,794	29,406	31,707	34,854	39,793	17,506	296,597
Trial Phase												
0	-	-	-	2	2	-	15	8	10	6	-	43
I	1,624	2,359	2,680	2,537	2,683	2,598	2,584	3,476	2,916	2,975	1,204	27,636
I and II	585	669	755	780	792	784	829	959	1,022	1,148	431	8,754
II	2,858	3,239	3,245	3,217	3,180	3,190	3,244	3,182	3,620	4,526	1,582	35,083
II and III	335	378	397	378	398	410	414	455	543	553	210	4,471
III	2,136	2,634	2,312	2,686	2,589	2,567	2,545	2,304	3,251	4,193	1,174	28,391
III and IV	11	12	22	20	29	14	44	43	41	24	14	274
IV	1,755	2,208	2,067	2,078	2,146	2,132	2,269	2,481	2,647	2,862	1,005	23,650
Not Specified	7,198	11,850	12,990	15,105	13,596	15,099	17,462	18,799	20,804	23,506	11,886	168,295
Per-trial Sample Size												
01 - 09	1,049	1,369	1,461	1,539	1,652	1,627	1,805	1,749	1,467	1,073	380	15,171
10 - 99	7,197	9,996	11,018	11,657	12,082	12,339	14,322	16,557	17,058	19,082	9,194	140,502
100 - 999	6,059	8,882	9,074	10,310	9,500	10,121	11,005	11,260	13,265	15,784	6,507	111,767
1000 - 9999	1,051	1,673	1,422	1,731	1,495	1,490	1,632	1,611	2,018	2,544	990	17,657
10000 - 99999	91	172	171	289	199	237	232	239	265	351	135	2,381
100000 or More	9	17	10	39	26	33	43	60	64	90	37	428
Not Specified	1,046	1,240	1,312	1,238	461	947	367	231	717	869	263	8,691
Geographical Region												
Northern America	13,383	17,007	17,149	17,742	18,239	19,661	20,515	23,509	24,221	27,527	12,027	210,980
Asia	452	1,023	1,978	2,975	3,633	2,711	5,031	6,223	5,759	6,050	3,892	39,727
Europe	2,214	4,880	4,651	5,579	2,799	3,840	2,388	552	3,191	4,449	742	35,285
Oceania	453	439	690	507	697	450	1,378	1,329	1,370	1,287	742	9,342
Latin America	0	0	0	0	47	132	94	94	313	480	103	1,263
Registry Name												
CT.gov	13,383	17,007	17,149	17,742	18,239	19,661	20,515	23,509	24,221	27,527	12,027	210,980
EUCTR	2,214	4,880	4,651	5,579	2,799	3,840	2,388	552	3,191	4,449	742	35,285
JPRN	370	639	1,327	1,927	2,277	775	2,908	3,525	2,798	2,344	2,099	20,989
CHICTR	61	279	332	439	685	1,139	1,138	1,643	1,839	2,567	1,036	11,158
ACTRN	453	439	690	507	697	450	1,378	1,329	1,370	1,287	742	9,342
CTRI	9	86	308	597	658	785	949	1,018	1,091	1,111	747	7,359
RBR	0	0	0	0	47	132	94	94	313	480	103	1,263
SLCTR	12	19	11	12	13	12	36	37	31	28	10	221

Notes: This extends Table 1 and details the number of registered clinical trials across categories. This table provides summary statistics of clinical trials registered in the WHO International Clinical Trials Registry Platform (ICTRP, `http://www.who.int/ictrp/en/`, retrieved in October 2017). The sample consists of clinical trials registered there between January 1st, 2007 to May 30th, 2017. See Section 2 for discussions about this exhibit and Appendix A.2.1 for the computational procedure.

Table A.3: Magnitude of a Part of the Clinical Trial Industry: Details 3

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2007-2017
Sum of Sample Sizes	13,965,001	21,737,552	15,076,229	29,863,810	28,395,906	31,615,304	34,383,227	38,990,898	61,796,652	72,113,892	19,964,109	367,902,580
Trial Phase												
0	0	0	0	80	110	0	1,792	464	538	1,928	0	4,912
I	127,469	175,775	179,879	220,763	319,846	204,398	207,803	223,460	187,132	200,013	90,116	2,136,654
I and II	37,341	47,223	62,080	48,263	56,722	59,693	65,346	67,594	69,814	99,567	30,130	643,773
II	327,128	519,669	469,622	443,725	350,115	380,366	347,675	337,626	405,371	572,493	183,170	4,336,960
II and III	83,047	77,214	93,265	163,681	59,655	73,923	106,499	116,128	141,267	160,358	46,398	1,121,435
III	1,334,347	1,594,702	1,491,894	1,670,677	1,573,894	2,725,446	1,962,131	1,466,568	2,239,443	3,357,147	821,578	20,237,827
III and IV	3,097	4,770	5,248	4,102	8,191	1,170	6,063	6,695	24,893	29,395	1,904	95,528
IV	911,295	1,683,270	681,033	620,357	1,050,000	1,168,265	1,014,552	1,079,280	1,597,887	1,312,236	457,628	11,575,803
Not Specified	11,141,277	17,634,929	12,093,208	26,692,162	24,977,373	27,002,043	30,671,366	35,693,083	57,130,307	66,380,755	18,333,185	327,749,688
Per-trial Sample Size												
01 - 09	3,662	4,643	4,695	4,828	5,160	4,818	5,633	5,641	4,638	3,703	1,798	49,219
10 - 99	305,354	424,348	466,263	499,924	513,324	535,295	617,967	709,634	749,169	846,317	405,297	6,072,892
100 - 999	1,875,067	2,843,986	2,781,065	3,179,257	2,818,809	3,046,945	3,128,180	3,019,024	3,646,255	4,406,802	1,708,506	32,453,896
1000 - 9999	2,821,333	3,752,555	3,264,929	4,264,126	3,640,623	3,620,286	3,844,087	3,901,555	4,770,421	6,111,351	2,346,775	42,338,041
10000 - 99999	2,024,433	3,765,116	3,422,652	5,856,220	5,008,243	5,579,831	5,314,855	5,739,832	6,369,719	9,341,660	3,000,031	55,422,592
100000 or More	6,935,152	10,946,904	5,136,625	16,059,455	16,409,747	18,828,129	21,472,505	25,615,212	46,256,450	51,404,059	12,501,702	231,565,940
Geographical Region												
Northern America	11,942,220	15,523,248	11,051,412	23,359,508	24,520,087	25,862,645	29,151,643	35,440,244	53,753,160	64,408,969	16,523,490	311,536,626
Asia	257,595	2,338,505	546,402	1,699,612	1,219,104	2,443,280	1,932,354	2,112,047	5,032,574	4,053,759	2,357,079	23,992,311
Europe	1,602,265	3,718,367	3,244,902	4,693,969	2,279,263	3,182,499	2,147,202	615,088	2,129,654	2,684,820	503,214	26,801,243
Oceania	162,921	157,432	233,513	110,721	373,721	119,235	1,144,156	815,080	830,148	663,391	570,083	5,180,400
Latin America	0	0	0	0	3,731	7,645	7,872	8,439	51,116	302,953	10,243	392,000
Registry Name												
CT.gov	11,942,220	15,523,248	11,051,412	23,359,508	24,520,087	25,862,645	29,151,643	35,440,244	53,753,160	64,408,969	16,523,490	311,536,626
EUCTR	1,602,265	3,718,367	3,244,902	4,693,969	2,279,263	3,182,499	2,147,202	615,088	2,129,654	2,684,820	503,214	26,801,243
JPRN	223,288	2,152,362	355,719	802,470	650,381	182,136	931,773	1,007,015	2,314,822	1,580,687	1,138,393	11,339,046
CHICTR	31,115	144,069	122,229	419,501	403,900	1,294,360	470,333	776,319	2,430,995	1,829,686	1,051,623	8,974,131
ACTRN	162,921	157,432	233,513	110,721	373,721	119,235	1,144,156	815,080	830,148	663,391	570,083	5,180,400
CTRI	920	13,936	64,767	476,352	161,508	964,550	473,406	321,941	280,283	634,959	165,221	3,557,842
RBR	0	0	0	0	3,731	7,645	7,872	8,439	51,116	302,953	10,243	392,000
SLCTR	2,272	28,138	3,687	1,289	3,315	2,234	56,842	6,772	6,474	8,427	1,842	121,292

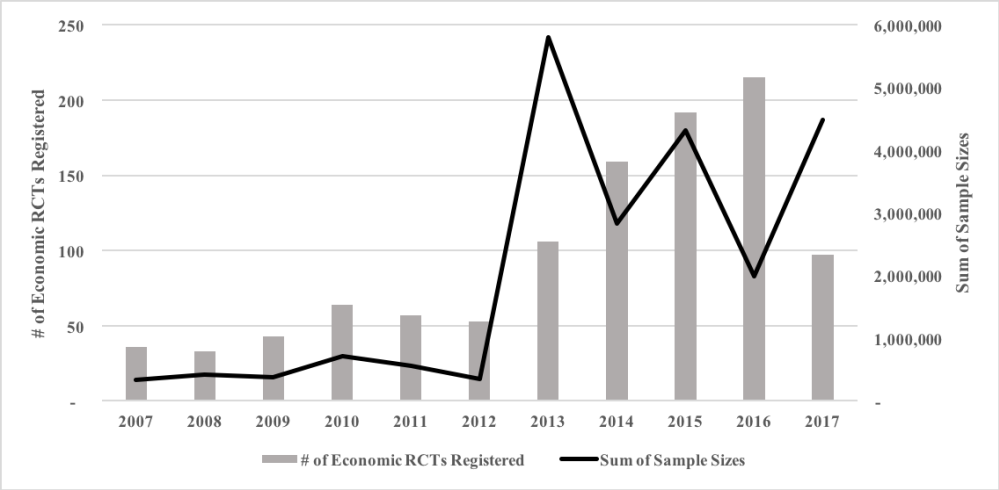
Notes: This extends Table 1 and details trial sample sizes across categories. This table provides summary statistics of clinical trials registered in the WHO International Clinical Trials Registry Platform (ICTRP, `http://www.who.int/ictrp/en/`, retrieved in October 2017). The sample consists of clinical trials registered there between January 1st, 2007 to May 30th, 2017. See Section 2 for discussions about this exhibit and Appendix A.2.1 for the computational procedure.

Table A.4: Magnitude of a Part of Economic RCTs

(a) Registered Economic RCTs & Sample Sizes

	Sample Period 2007-2017 May
Total Number of Economic RCTs Registered	1055
Sum of Sample Sizes	22,190,304

(b) Time Evolution



Notes: This table provides summary statistics of economic RCTs registered in the American Economic Association RCT Registry (`https://www.socialscienceregistry.org`, retrieved in October 2017). The sample consists of RCTs registered there between January 1st 2007 to May 30th 2017 and where the unit of outcome measurement is an individual or a household. I focus on RCTs with individual or household subjects in order to make it possible to sum up sample sizes. See Section 2 for discussions about this exhibit and Appendix A.2.1 for the detailed computational procedure.

Table A.5: Magnitude of a Part of Economic RCTs: Details 1

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2007-2017
# of Economic RCTs Registered	36	33	43	64	57	53	106	159	192	215	97	1,055
Sum of Sample Sizes	333,020	414,928	380,036	715,767	562,845	347,793	5,804,526	2,836,986	4,321,824	1,987,181	4,485,398	22,190,304
Per-trial Sample Size												
1-99	-	-	-	2	1	2	3	5	12	7	1	33
100-999	6	7	8	15	9	11	30	36	62	78	37	299
1000-9999	22	21	26	39	32	36	59	92	87	99	44	557
10000-99999	8	4	8	5	13	3	12	24	27	26	14	144
100000-999999	-	1	1	3	1	1	1	1	3	5	-	17
1000000 or More	-	-	-	-	-	-	1	1	1	-	1	4
Not specified	-	-	-	-	-	-	-	-	-	-	-	-
Geographical Region												
North America	8	9	4	10	10	6	18	28	36	34	16	179
South America	7	2	9	10	6	11	8	15	18	29	4	119
Asia	9	11	6	18	13	16	28	45	45	45	18	254
Africa	9	7	19	19	24	14	43	45	59	49	10	298
Europe	3	4	5	7	2	5	5	14	23	29	19	116
Others/Not specified	-	-	-	-	-	-	-	-	-	-	-	-
Keywords												
Agriculture	3	1	6	8	8	2	5	22	12	18	9	94
Education	13	13	11	17	14	17	34	49	45	57	18	288
Electoral	1	-	-	2	3	4	5	4	4	13	4	40
Environment & Energy	-	4	-	3	3	5	4	8	24	11	7	69
Finance & Microfinance	15	11	17	26	13	11	21	33	33	22	8	210
Governance	5	2	5	6	6	6	8	16	29	32	18	133
Health	14	9	9	11	23	14	36	43	59	44	22	284
Labor	4	4	10	12	8	9	25	26	32	64	23	217
Post-conflict	1	2	3	-	3	3	5	3	2	3	2	27
Welfare	5	3	2	8	9	4	13	18	35	38	20	155
Randomization Unit												
Area	3	4	2	3	2	2	5	5	14	8	6	54
Firm	1	3	4	3	6	1	9	6	10	14	3	60
Individual	17	10	19	31	18	27	53	81	107	112	67	542
School	3	3	4	5	6	8	11	14	15	15	4	88
Facility	1	-	2	2	2	1	2	4	3	10	2	29
Household	4	2	2	7	5	3	5	10	11	9	2	60
Village	6	5	5	5	12	4	15	16	16	17	2	103
Other	1	6	5	8	6	7	6	23	16	30	11	119

Notes: This extends Table A.4 and details the number of registered economic RCTs and sample sizes across categories. This table provides summary statistics of economic RCTs registered in the AEA RCT Registry (`https://www.socialscienceregistry.org`, retrieved in October 2017). The sample consists of RCTs registered there between January 1st 2007 to May 30th 2017 and where the unit of outcome measurement is an individual or a household. "Randomization Unit" is the unit of treatment assignment, which may be different from the unit of outcome measurement. A single RCT may have multiple "Keywords." See Section 2 for discussions about this exhibit and Appendix A.2.1 for the computational procedure.

Table A.6: Magnitude of a Part of Economic RCTs: Details 2

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2007-2017
# of Economic RCTs Registered	-	-	1	4	8	4	8	13	12	14	7	71
Sum of Sample Sizes	-	-	210	3,841	10,356	5,235	9,962	28,006	5,377	8,587	4,808	76,382
Per-trial Sample Size												
1-99	-	-	-	1	2	-	1	-	2	-	1	7
100-999	-	-	1	2	3	2	3	6	5	3	3	28
1000-9999	-	-	-	1	2	2	3	3	2	5	2	20
10000-99999	-	-	-	-	-	-	-	1	-	-	-	1
100000-999999	-	-	-	-	-	-	-	-	-	-	-	-
1000000 or More	-	-	-	-	-	-	-	-	-	-	-	-
Not specified	-	-	-	-	1	-	1	3	3	6	1	1
Geographical Region												
North America	-	-	-	-	-	-	2	1	1	1	1	6
South America	-	-	1	1	2	-	1	3	-	3	1	12
Asia	-	-	-	1	4	2	2	4	4	2	2	21
Africa	-	-	-	2	2	2	2	4	3	3	1	19
Europe	-	-	-	-	-	-	1	1	2	3	1	8
Others/Not specified	-	-	-	-	-	-	-	-	-	-	-	-
Keywords												
Agriculture	-	-	-	1	-	-	-	1	-	-	-	2
Education	-	-	1	-	2	-	4	3	1	2	1	14
Electoral	-	-	-	-	-	-	-	-	1	-	-	1
Environment & Energy	-	-	-	1	-	-	-	-	-	2	2	5
Finance & Microfinance	-	-	-	-	1	1	2	4	3	3	1	15
Governance	-	-	-	-	4	2	-	3	3	2	3	17
Health	-	-	-	2	4	-	3	1	3	2	-	15
Labor	-	-	-	-	3	1	2	1	3	3	5	18
Post-conflict	-	-	-	-	-	-	-	1	-	1	1	3
Welfare	-	-	-	-	1	1	4	1	3	4	-	14
Randomization Unit												
Area	-	-	-	-	-	2	-	-	-	1	2	5
Firm	-	-	-	1	3	2	2	6	4	5	3	26
Individual	-	-	-	-	-	-	-	-	-	-	-	-
School	-	-	-	-	1	-	2	2	1	-	-	6
Facility	-	-	-	2	3	-	2	-	1	-	-	8
Household	-	-	-	-	-	-	-	-	-	-	-	-
Village	-	-	-	1	-	-	1	2	1	-	-	5
Other	-	-	1	-	-	-	-	-	2	4	1	8

Notes: This extends Table A.4 and details the number of registered registered economic RCTs and sample sizes across categories. This table provides summary statistics of economic RCTs registered in the AEA RCT Registry (`https://www.socialscienceregistry.org`, retrieved in October 2017). The sample consists of RCTs registered there between January 1st 2007 to May 30th 2017 and where the unit of outcome measurement is *not* an individual or a household. "Randomization Unit" is the unit of treatment assignment, which may be different from the unit of outcome measurement. A single RCT may have multiple "Keywords." See Section 2 for discussions about this exhibit and Appendix A.2.1 for the computational procedure.

Table A.7: A Selection of High-stakes RCTs (Continued from Table 2)

(a) Medical Clinical Trials

	Subjects	Sample Size
i	Coronary Heart Disease Patients	4444 Individuals
ii	Hypertensive Patients with Diabetes	1148 Individuals
iii	Patients with Elevated Intraocular Pressure	1636 Individuals
iv	HIV Negative Gay Men and Transgender Women	2499 Individuals
v	Serodiscordant Couples	1763 Couples
vi	Postmenopausal Women	16608 Individuals

(b) Social and Economic Experiments

	Subjects	Sample Size
I	Poor Households in Kenya	940 Households
II	Crime Hot Spots in Minneapolis	110 Spots
III	Unmarried Women in Malawi	1007 Individuals
IV	Uninsured Individuals in Oregon	12229 Individuals
V	Public Sector Job Applicants in Mexico	350 Job Vacancies

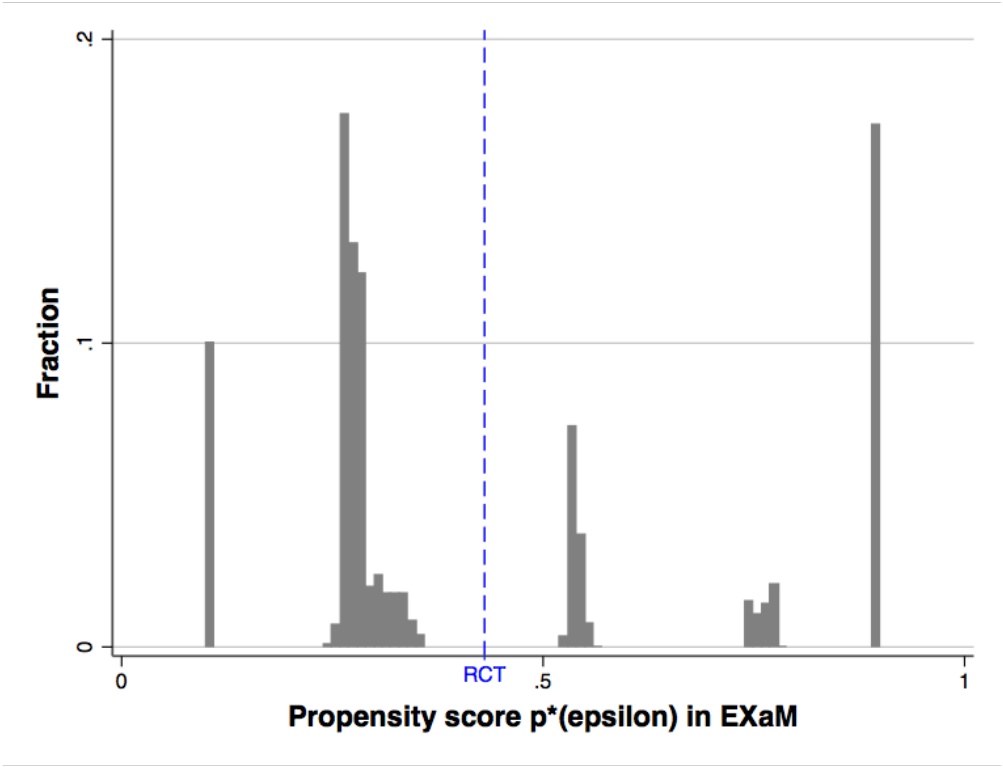
Notes: This is a continuation of Table 2. See Table 2's notes about how to read this table.

Table A.8: Summary Statistics of Treatment Effects and WTP

	(1) Mean	(2) 25 Percentile	(3) Median	(4) 75 Percentile	(5) SD
% Reduction in Child diarrhea	4.89	3.49	5.00	6.66	3.52
WTP measured by time cost of water collection (unit: workdays)	24.99	-6.20	19.22	46.49	137.4

Notes: This table shows summary statistics of estimated treatment effects $\hat{e}_{t_1 i}$ and WTP $\hat{w}_{it_1}$. I bootstrap $\hat{e}_{t_1 i}$ and $\hat{w}_{it_1}$ from their estimated models (10) and (11), respectively. See Section 7.2 for discussions about this table.

Figure A.1: EXaM's Treatment Assignment Probabilities



Notes: This figure shows the distribution of EXaM's treatment assignment probabilities $p^*_{it_1}(\epsilon)$ over 1000 bootstrap simulations and households. The vertical dash line is RCT's constant assignment probability $p^{RCT}_{it_1}$. See Section 7.3 for discussions about this table.