[NOT FOR PUBLICATION]

Web Appendix to Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes

James J. Heckman,[1] Rodrigo Pinto, and Peter A. Savelyev

The University of Chicago

November 26, 2012

[1]James Heckman is the Henry Schultz Distinguished Service Professor of Economics and Public Policy at the University of Chicago; Professor of Science and Society, University College Dublin; and Senior Fellow at the American Bar Foundation. Rodrigo Pinto is a Ph.D. Candidate in Economics at the University of Chicago. Peter Savelyev is an Assistant Professor of Economics at Vanderbilt University and a Health Policy Associate of the Robert Wood Johnson Center for Health Policy at Meharry Medical College. We thank the editor, Robert Moffit, and three anonymous referees for helpful comments. A version of this paper was presented at a seminar at the HighScope Foundation, Ypsilanti, MI, December 2006; at a conference at the Minneapolis Federal Reserve, Minneapolis, MN, December 2007; at a National Poverty Center conference, Ann Arbor, MI, December 2007; at a conference sponsored by the Jacobs Foundation at Castle Marbach, Germany, April 2008; at the Leibniz Network Conference on noncognitive skills, Mannheim, Germany, May 2008; at an Institute for Research on Poverty conference, Madison, WI, June 2008; at the Society for Research on Child Development, Denver, CO, April 2009; at the Association for Research in Personality Conference, Evanston, IL, July 2009; at the Public Policy & Economics Workshop at the Harris School of Public Policy at the University of Chicago, Chicago, IL, October 2009, which was attended by Diane Schatzenbach; at the Cultivating Human Capital Conference, Chicago, IL, December 2009; at an IFS seminar at University College London, London, England, April 2010; at the Brookings Institution, Center for Universal Education, Washington DC, June 2010; at the NBER Summer Institute, Economics of Crime working group, Cambridge, MA, July 2010; and at an Applied Economics Luncheon, Booth School, Chicago, May 2011. We thank participants at these meetings for useful comments. We are grateful to Clancy Blair, Dan Benjamin, Martin Browning, Sarah Cattan, Kenneth Dodge, Angela Duckworth, Amy Finklestein, Miriam Gensowski, Matt Gentzkow, Jeff Grogger, Emir Kamenica, Costas Meghir, Jörn-Steffen Pischke, Devesh Raval, Brent Roberts, Tino Sanandaji, Larry Schweinhart, Sandra Waxman, Ben Williams, and Junjian Yi for helpful comments. We are grateful to Christopher Hansman, Kegon Tan Teng Kok, Min Ju Lee, Xiliang Lin, Yun Pei, and Ivana Stosic for excellent research assistance. This research was supported in part by the American Bar Foundation, the JB & MK Pritzker Family Foundation, Susan Thompson Buffett Foundation, NICHD R37HD065072, R01HD54702, a grant to the Becker Friedman Institute for Research and Economics from the Institute for New Economic Thinking (INET), and an anonymous funder. We acknowledge the support of a European Research Council grant hosted by University College Dublin, DEVHEALTH 269874. We thank the HighScope Foundation for supplying us with the data used in this paper. The views expressed in this paper are those of the authors and not necessarily those of the funders or commentators mentioned here. Supplementary materials are placed in a Web Appendix.

Contents

List of Tables

List of Figures

A The Perry Preschool Program Curriculum

The HighScope Perry Preschool program (called the Perry program in the text) was an early childhood educational experiment conducted in Ypsilanti, Michigan during the early 1960s. The study enrolled five annual entry cohorts during the period 1961–1965, totaling 123 children (58 treatment and 65 control). Children were admitted at age three for a two-year program, except for those of the first cohort, who were admitted at age four and received only one year of the program. The last wave was taught alongside a group of three-year-olds who were not included in the Perry analysis sample. Drawn from the African-American population surrounding the Perry Elementary School, subjects were located through a survey of families associated with the school, as well as through neighborhood group referrals and door-to-door canvassing. Disadvantaged children were identified by entry IQ and an index of socioeconomic status (SES). Those with IQ scores[1] outside the range of 70–85 were excluded,[2] as were those with organic mental defects. SES was measured using a weighted linear combination of three components: paternal employment skill level, parental educational attainment, and the number of rooms per person in the family home. Subjects with SES above a certain level (fixed at study inception) were excluded. The average yearly program cost was $9,825 per participant in U.S. CPI-adjusted 2006 dollars (Heckman et al., 2010*a*, Table C.1 of the Web Appendix to that paper). Multiple measurements on outcomes were taken at ages 3–15, 19, 27 and 40.

Preschool Overview Each preschool class had 20–25 children. The program consisted of 2.5-hour preschool classes on weekdays during the school year (30 weeks per year, October through May), supplemented by weekly 1.5-hour home visits by teachers. Teachers had special training for tutoring disadvantaged children and were certified for elementary, early childhood, and special education. The child-teacher ratio ranged from 5 to 6.25 over the course of the program (Schweinhart, Barnes and Weikart, 1993, p.32).

[1] Measured by the Stanford-Binet IQ test (1960s norming, see Appendix B).

[2] Compromises in selection and randomization protocols are discussed by Heckman et al. (2010*b*)

Home Visits Weekly home visits, each lasting 1.5 hours, were conducted by the preschool teachers. The purpose of these visits was to "involve the mother in the educational process" and "implement the curriculum at home" (Schweinhart, Barnes and Weikart, 1993, p.32). During the visit, teachers encouraged mothers to participate in their child's education and helped with any problems arising in the home. Occasionally, these visits took the form of field trips to stimulating environments such as the zoo.

Curriculum The Perry curriculum was based on the principle of *active participatory learning*, in which children and adults are seen as equal partners in the learning process.[3] In active participatory learning, children are engaged with objects, people, events, and ideas. Children's abilities to plan, execute, and evaluate tasks are fostered, as are their social skills, including cooperation with others and resolution of interpersonal conflicts. The curriculum of the Perry program was grounded in the research on cognitive development by Piaget and Inhelder (2000), the progressive educational philosophy of Dewey (1997), and the sociocultural theories of Vygotsky (1986). The signature of the curriculum was the *plan-do-review sequence* in which children actively made choices about what they would do, purposefully carried out their ideas, and reflected on their activities and what they learned. Children also engaged in small- and large-group activities, initiated by teachers, which encouraged their independent use of classroom materials and investigation of ideas. Activities followed a consistent daily routine. The classroom was well supplied with diverse learning materials organized and labeled to help children find, use, and return the materials on their own. The curriculum's educational content was organized around key experiences (called "key developmental indicators") that help to develop skills in language and literacy, mathematics, initiative and social relations, and the arts. Teachers assisted children's learning in these key areas by asking open-ended questions (e.g., "Can you show me how you made that?") and encouraging independent problem solving (e.g., "How can you get it to fit?"). For a com-

[3]The curriculum is described in Schweinhart, Barnes and Weikart 1993, pp.34–36; and Weikart, Bond and McNeil 1978, pp.21–23.

plete description of the curriculum's content and teaching practices, see Hohmann, Weikart and Epstein (2008). Features such as the plan-do-review sequence, room arrangement, and a structured daily routine were intended to help children "develop a sense of responsibility and to enjoy opportunities for independence" (Schweinhart, Barnes and Weikart, 1993, p.32-33). The Perry curriculum has been interpreted as implementing the Vygotskian principles currently advocated in Tools of the Mind. (See Sylva, 1997, and Bodrova and Leong, 2001.)

B Cognitive Tests

B.1 Stanford-Binet

The Stanford-Binet Intelligence Scale (Terman and Merrill, 1960) is a revision of an earlier version of the test (Terman and Merrill, 1937) and is a measure of general intelligence.[4] The history of the Stanford-Binet test is presented in Becker (2003). In the 1950s, Merrill revised the Stanford-Binet by selecting the best items from Forms L (for Lewis) and M (for Maud)[5] from the 1937 version of the test. These were combined to create the Form L-M, published in 1960. The L-M form added alternate items at all levels, but otherwise remained similar in format to the 1937 forms.

The 1960 version of the Stanford-Binet is widely preferred over the 1937 version. In addition to retaining the best items of forms L and M, the alternative items added at each age level improved the accuracy of the test. The 1937 version of the test did not have a uniform standard deviation, and the IQs obtained from that version were not comparable across ages (Becker, 2003). The 1960 revision normalized the standard score to a mean of 100 and a standard deviation of 16 for all age groups.

Despite these improvements, the 1960 version of the Stanford-Binet has its limitations. The test has a ceiling, the maximum score that an examinee can get. According to a study by Kennedy et al. (1960) that surveyed mathematically gifted adolescents in a Summer Mathematical Institute at Florida State University, this ceiling makes the test inadequate when examining gifted adolescents. However, this is not a concern when analyzing the Perry program, as participants were selected to have low Stanford-Binet IQ. In addition to its inadequacy with respect to gifted children, the test has also been criticized for being dependent on language skills (Wade, 1978).

Although the 1960 version of the Stanford-Binet measures one general factor (general

[4] A single age scale is a test which is different in each age. It was used to provide a direct translation of each child's performance to his/her mental age (Becker, 2003).

[5] Lewis is the first name of Terman, and Maud is the first name of Merrill (Becker, 2003)

intelligence), Sattler's classification schema (Sattler, 1965) classifies one's performance on the test into seven major categories. A short description of the each major category follows.

Language: This category includes items which deal with maturity of vocabulary. It measures the number of words the subject can define and the extent of the subject's comprehension of verbal relations.

Memory: This category includes all the items which could be subclassified into meaningful memory (short stories), nonmeaningful memory (words), and visual memory.

Conceptual Thinking: This category, while closely associated with language ability, is primarily concerned with abstract thinking.

Reasoning: This category is subclassified into verbal and nonverbal reasoning. Reasoning includes the perception of logical relations, discrimination ability (understanding differences), analysis, and synthesis. A spatial reasoning factor is also included in the orientation items.

Numerical Reasoning: This category includes items specifically geared to numerical or arithmetical problems. Numerical reasoning includes such factors as concentration and the ability to generalize from numerical data.

Visual Motor: This category contains items concerned with manual dexterity, hand-eye coordination, and perception of spatial relations. Constructive visual imagery may be involved in such items as paper folding. This area is closely associated with nonverbal reasoning.

Social Intelligence: This category overlaps a good deal with the reasoning category. Social intelligence includes aspects of social maturity and social judgement.

Terman and Merrill (1960) present evidence on the reliability of the Stanford-Binet Scale using biserial correlations.[6] According to the manual, average biserial correlations tend to be highest at the adult levels, ranging from a low of .64 to a high of .80. At the preschool level the average biserial correlations are lower, with the lowest average of 0.53 obtained at age three.[7]

[6]The biserial correlation coefficient is a statistic that is used to indicate the strength of the relationship between a single item and the score on a test that includes the item. The correlation shows the extent to which the question is measuring the same knowledge or skill that the total test is measuring (Glass and Hopkins, 1995).

[7]Some of these results are summarized by Himelstein (1966).

B.2 Leiter

The Arthur Adaption of the Leiter International Performance Scale[8] is a test of nonverbal intelligence for young children. The test is given individually. The testing materials consist of frames, each with a sliding metal clip, several blocks, and several pattern strips. The sliding metal clip is used to hold the pattern strips. The subject must place the blocks in the proper position at stalls based on the information given by the pattern strips. If the subject successfully passes one stage, a frame for the next level is presented.

Scoring for the Arthur Adaptation of the Leiter scale follows the general principle of the mental age scale. The examination is begun at a level two years below the chronological age of the subject if the child appears to be of average intelligence. For instance, a five-year old child of apparently normal ability would first be given the age-three test. If the child passes the age-three test, the child would have a basal mental age of three. However, if the child fails the test, it would be necessary to go down to the age-two level to obtain a basal year. After the basal year is established, all the tests above the basal year are presented, including tests which the subject previously failed. Each test passed at the age-four level earns an additional two and a half months of mental age beyond the basal mental age. Each test passed at the age-twelve level earns an additional six months of mental age. Every other test passed beyond the basal mental age earns an additional three months of mental age. The testing is continued until the subject fails all the tests at two successive year levels. This increases the accuracy of the test by minimizing the effects of accidental failures. If a child shows apparent signs of mental retardation, the initial examination level is chosen at two years below his estimated mental age.

In the literature, the Arthur Adaptation of the Leiter scale is considered to have three major advantages over verbal tests: First, it allows for testing of children at lower chronological age levels than other performance scales since children can understand the directions of the test without any verbal explanation. Second, it tests the ability to learn rather than

[8]The information provided below is based on Arthur (1952).

early acquired skills or knowledge. For the first five tests, the subject is given credit as having passed if he/she is able to perform the task without demonstration or help during any one trial, irregardless of the number of previous trials that have been given and the level of demonstration or help given during previous trials. Therefore, examinees can use what they learned in previous tests to complete the current test. Third, every test is given without a time limit, which allows the examinee to complete the test without time pressure. The idea behind unlimited time testing is to create opportunities for small children to demonstrate their true level of intelligence, since children with different personality patterns may react differently to the test. For instance, a more talkative child may take more time to complete the test, but that does not necessarily mean that the child's intelligence is lower. The Leiter test generally takes hours to complete. The examinee is therefore under observation for a long period under controlled conditions.

The Arthur Adaptation of the Leiter scale is re-standardized using middle-class Americans as the base. The term "middle-class" is used on the basis of the occupational classification of the parents. Few cases were found at either extreme of the occupational scale, with the mass of the cases in the skilled and semi-skilled labor groups. The average population Leiter score is normalized to 100.

B.3 PPVT

The Peabody Picture Vocabulary Test (PPVT) (Dunn, 1965) is an assessment of verbal intelligence through the measurement of hearing vocabulary. Like the Leiter, the PPVT is administered individually.

As described in Dunn (1965), the PPVT consists of 150 plates, each with four numbered pictures. The pictures on each plate represent activities, objects, or states of being[9]. A plate with four pictures is first presented to the subject. Next, the examiner orally presents a stimulus word to the subject, and the subject is required to indicate which of the four

[9]'States of being' consist of a person's condition, attributes, personality, etc. For instance, a picture of a facial expression is a 'state of being' picture.

pictures on the plate best illustrates the meaning of the stimulus word. Credits are earned by correct indications. The difficulty of the plates increases over the course of the test. The subject earns a lowest estimate (basal) with eight consecutive correct answers and reaches a ceiling estimate with six errors on eight consecutive responses.

After completion of the test (attainment of the ceiling), a raw score is given based on the performance of the subject. The PPVT score is normalized to a mean of 100 and standard deviation of 15.

The advantages of the PPVT are its short testing time and the simplicity of the administration procedure. Brown and Rice (1967) and Taylor (1975) consider its narrow focus on verbal intelligence to be a disadvantage.

B.4 California Achievement Test

The California Achievement Test (CAT) (Tiegs and Clark, 1971) is used to assess the academic achievement of children. The CAT consists of three parts: reading (reading vocabulary and reading comprehension), arithmetic (arithmetic concepts and arithmetic problems), and language (language mechanics, language usage and structure, and spelling).

The CAT score reports provide the total score, the standardized score, the percentile score, and scores for each of the main parts and their subsections. In this paper, the total CAT score is used as an indicator of overall academic achievement.

Analysts conceptually distinguish the three IQ tests described above from the CAT test, since achievement exams test acquired skills and knowledge rather than pure intelligence. Furthermore, achievement tests are known to be loaded on social skills and highly loaded on general intelligence (Borghans et al., 2008, 2011). We therefore consider the IQ tests as measures of cognition, while we treat the achievement test as an outcome loaded on both cognitive and personality skills.

B.5 Relationships Between Different Measures of Cognition

Table B.1 compares correlations among scores from the Stanford-Binet, Leiter, PPVT, and CAT tests for the Perry sample. As shown in Table B.1, correlations between the Stanford-Binet and the other measures are above 0.6, while correlations between the Leiter and the PPVT are in the range of 0.25-0.42. The most likely reason for this substantial difference in correlations is that the Stanford-Binet IQ measures both verbal and non-verbal intelligence, while the Leiter is a measure of nonverbal intelligence and the PPVT is a measure of verbal intelligence. We also see in Table B.1 that IQ as measured by the PPVT is the least correlated with CAT performance. In figures B.1–B.4, we present empirical CDFs of Stanford-Binet, Leiter, PPVT and CAT scores at ages 7, 8, and 9. Figure B.5 shows density histograms of CAT total score at age 14. We show p-values for a difference in means test above each chart. The figures confirm that treatment raises IQ for females (with the exception of PPVT, see Figures B.3 and B.3), but not for males. Similar to IQ scores, CAT scores at ages 7, 8, and 9 are statistically significant for females, but not for males.

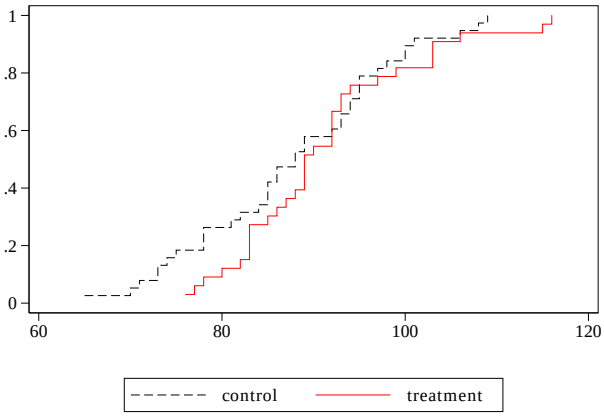
Table B.1: Correlations Among Stanford-Binet, Leiter, PPVT and CAT Scores in the Perry Sample

		Males				Females			
		Binet	Leiter	PPVT	CAT	Binet	Leiter	PPVT	CAT
Binet	correlation	1				1			
	p -value								
	N								
Leiter	correlation	.635 ***	1			.669 ***	1		
	p -value	*(.000)*				*(.000)*			
	N	72				51			
PPVT	correlation	.712 ***	.250 ***	1		.626 ***	.423 ***	1	
	p -value	*(.000)*	*(.034)*			*(.000)*	*(.002)*		
	N	72	72			51	51		
CAT	correlation	.662 ***	.648 ***	0.3539 ***	1	.713 ***	.619 ***	0.4331 ***	1
	p -value	*(.000)*	*(.000)*	*(.003)*		*(.000)*	*(.000)*	*(.002)*	
	N	71	71	71		50	50	50	

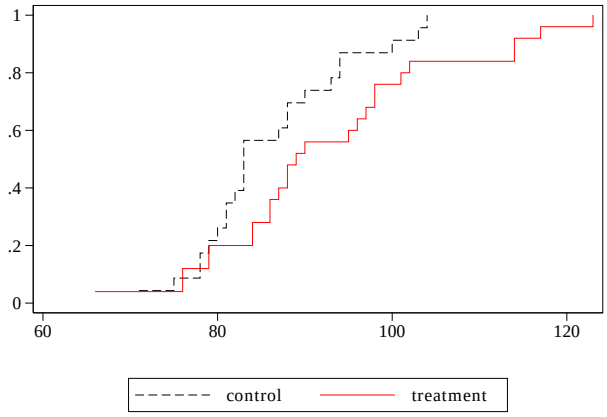
Notes: Pearson correlations are shown. N denotes the sample size. Each IQ score is an average over non-missing observations at ages 7, 8, and 9. Scores are from the Stanford-Binet Intelligence Scale (Terman and Merrill, 1960), the Leiter International Performance Scale (Arthur, 1952), the Peabody Picture Vocabulary Test (Dunn, 1965), and the California Achievement Test (Tiegs and Clark, 1971). Stars denote: *** - 1% significance level.

Figure B.1: Empirical CDFs of the Stanford-Binet Measures, Perry Sample

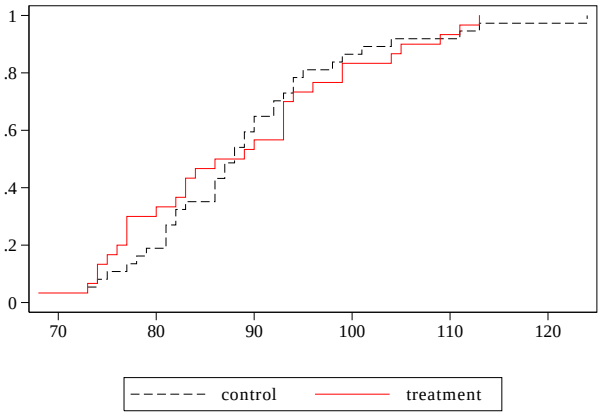
(a) Stanford-Binet Age 7
Males
$p = .088$;



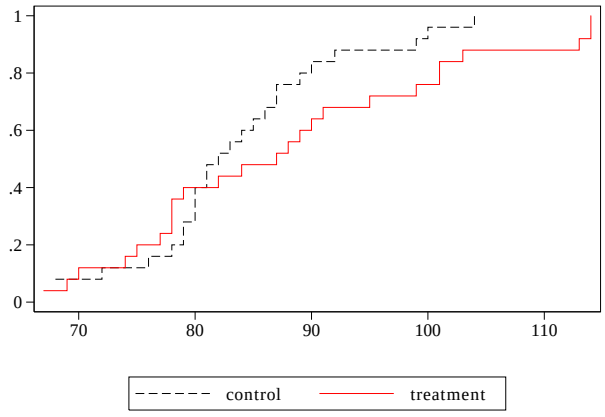
(b) Stanford-Binet Age 7
Females
$p = .030$



(c) Stanford-Binet Age 8
Males
$p = .598$



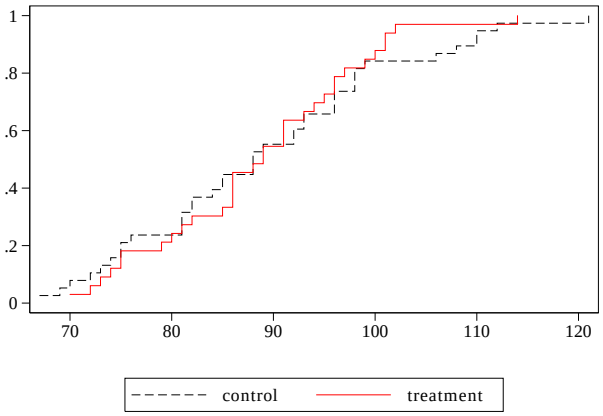
(d) Stanford-Binet Age 8
Females
$p = .105$



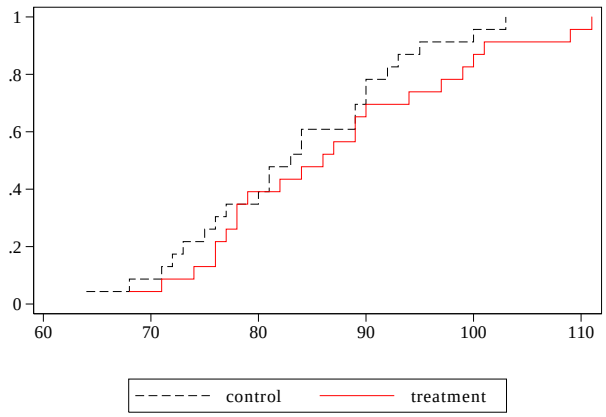
Notes: "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure B.1: Continued Empirical CDFs of the Stanford-Binet Measures, Perry Sample

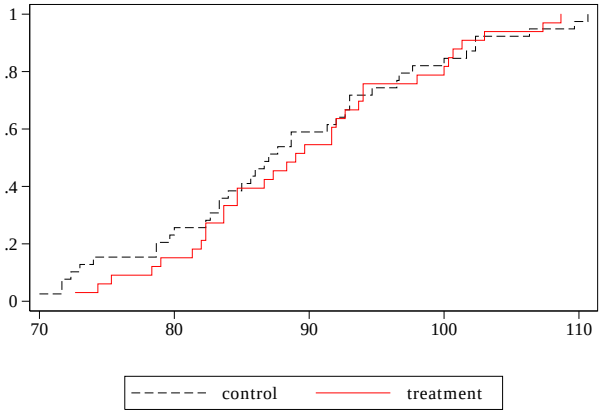
(e) Stanford-Binet Age 9
Males
$p = .587$



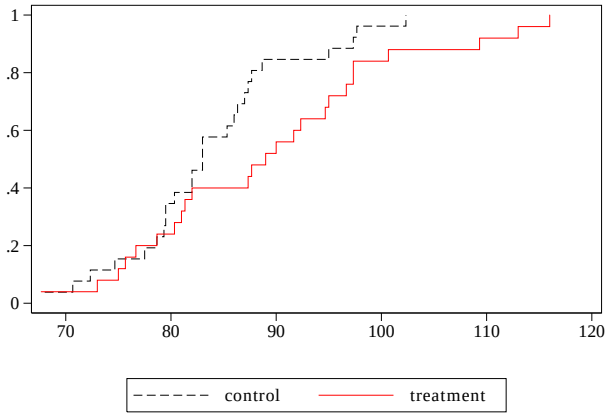
(f) Stanford-Binet Age 9
Females
$p = .133$



(g) Stanford-Binet Index[(a)]
Males
$p = .297$



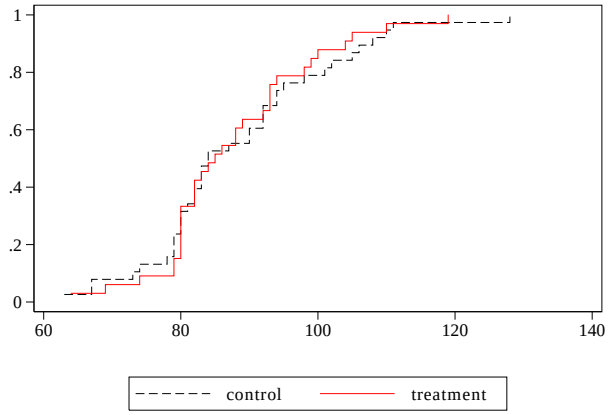
(h) Stanford-Binet Index[(a)]
Females
$p = .033$



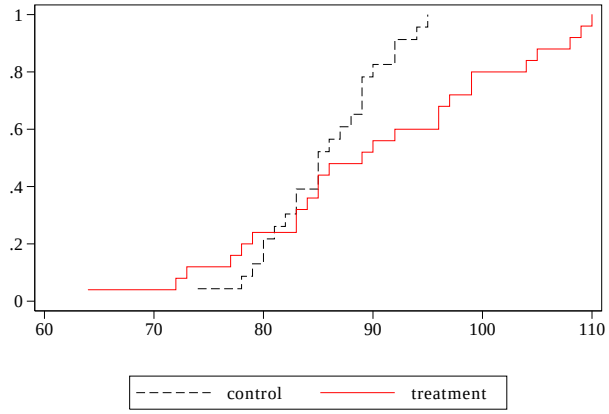
Notes: [(a)]The index is an average over ages 7, 8, and 9. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure B.2: Empirical CDFs of the Leiter Measures, Perry Sample

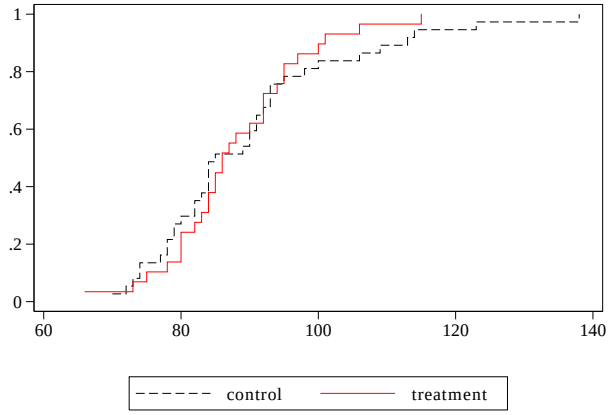
(a) Leiter Age 7, Males
$p = .580$



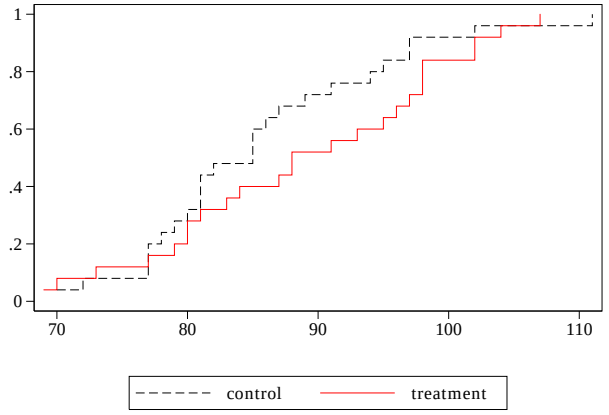
(b) Leiter Age 7, Females
$p = .069$



(c) Leiter Age 8, Males
$p = .706$



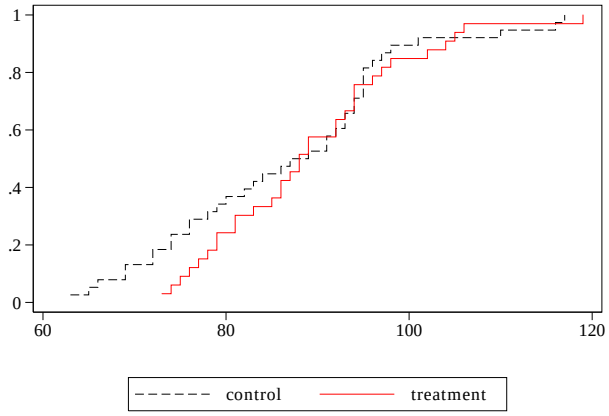
(d) Leiter Age 8, Females
$p = .136$



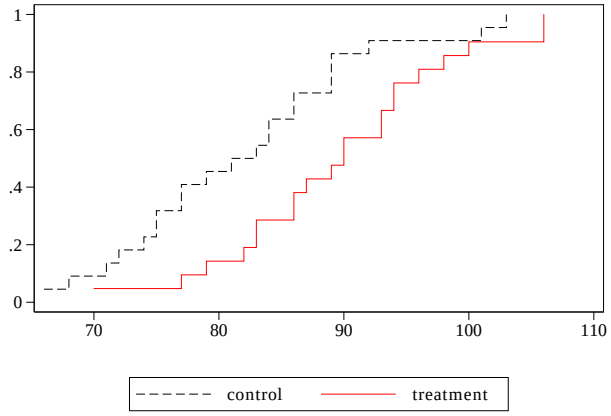
Notes: "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure B.2: Continued Empirical CDFs of the Leiter Measures, Perry Sample

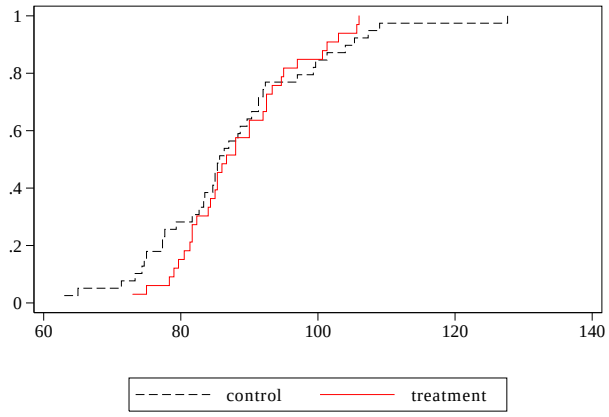
(e) Leiter Age 9, Males
$p = .188$



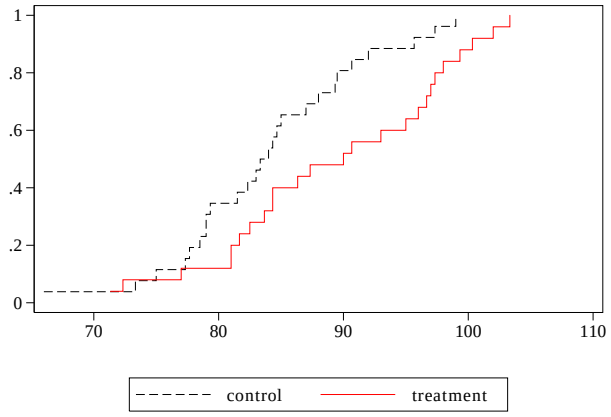
(f) Leiter Age 9, Females
$p = .005$



(g) Leiter Index, Males[(a)]
$p = .362$



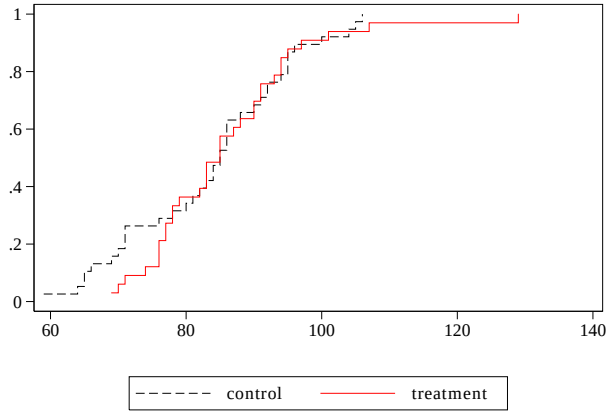
(h) Leiter Index, Females[(a)]
$p = .014$



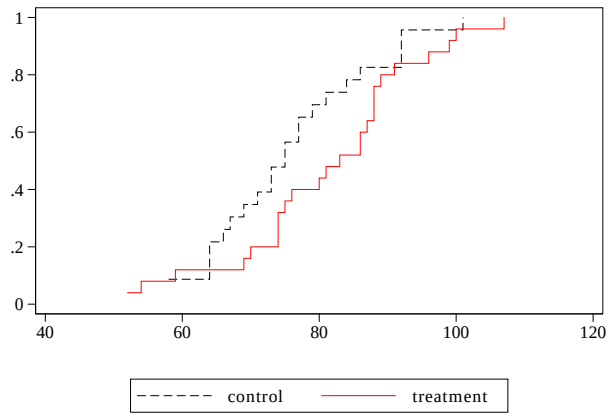
Notes: [(a)]The index is an average over ages 7, 8, and 9. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure B.3: Empirical CDFs of the PPVT Measures, Perry Sample

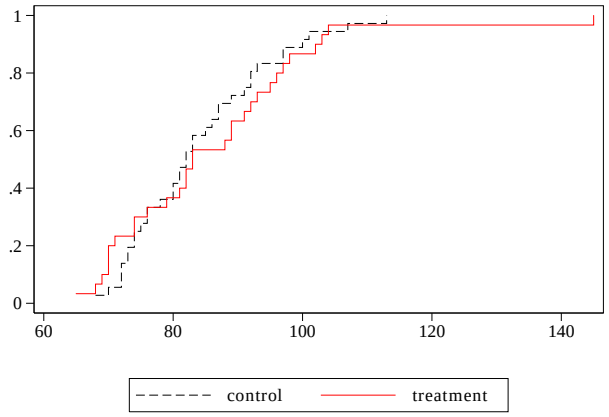
(e) PPVT Age 7, Males
$p = .214$



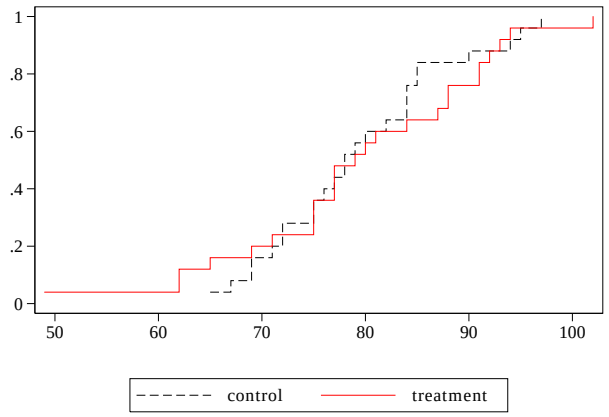
(f) PPVT Age 7, Females
$p = .072$



(g) PPVT Age 8, Males
$p = .273$



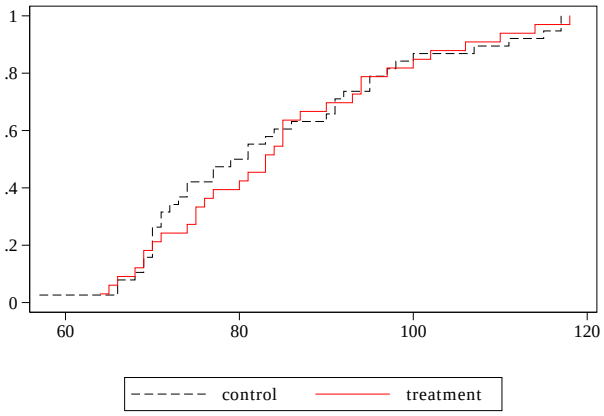
(h) PPVT Age 8, Females
$p = .495$



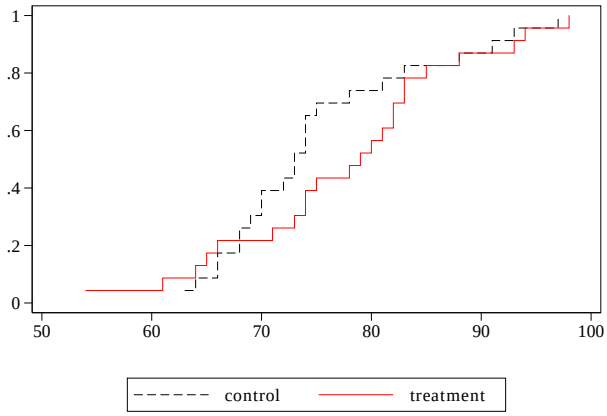
Notes: "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure B.3: Continued Empirical CDFs of the PPVT Measures, Perry Sample

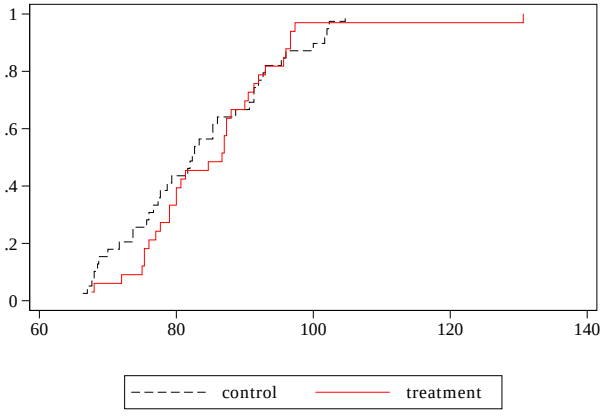
(i) PPVT Age 9, Males
$p = .390$



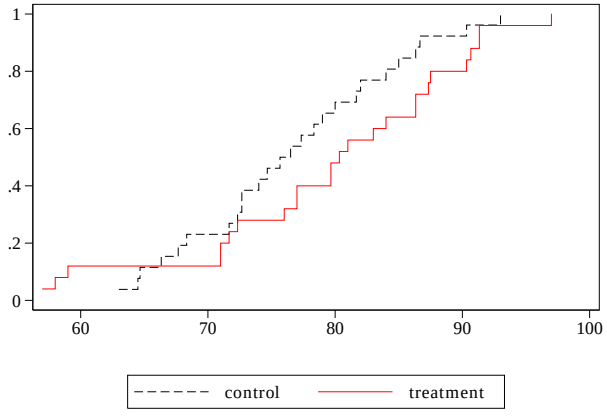
(j) PPVT Age 9, Females
$p = .224$



(k) PPVT Index, Males[a]
$p = .204$



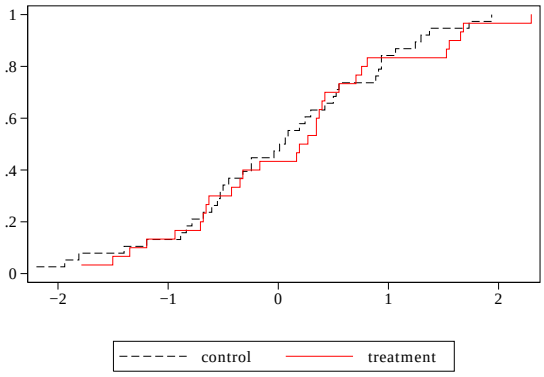
(k) PPVT Index, Females[a]
$p = .136$



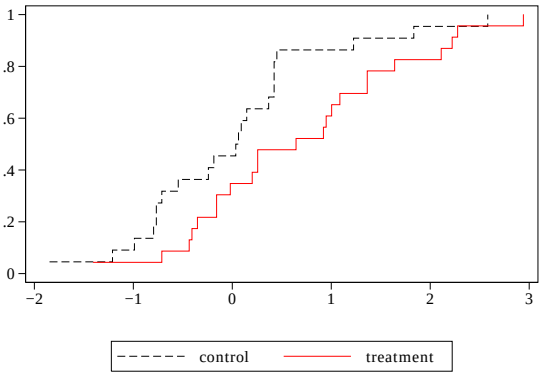
Notes: [a]The index is an average over ages 7, 8, and 9. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure B.4: Empirical CDFs of the CAT Measures

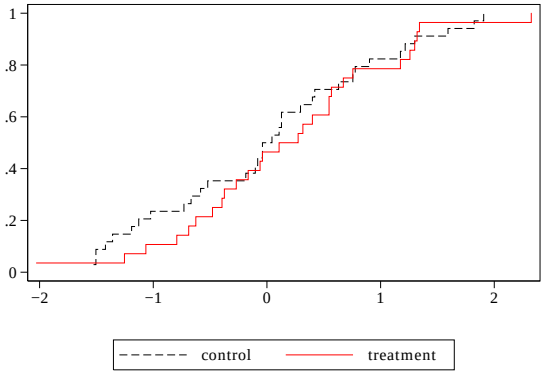
(a) CAT Age 7, Males
$p = .326$



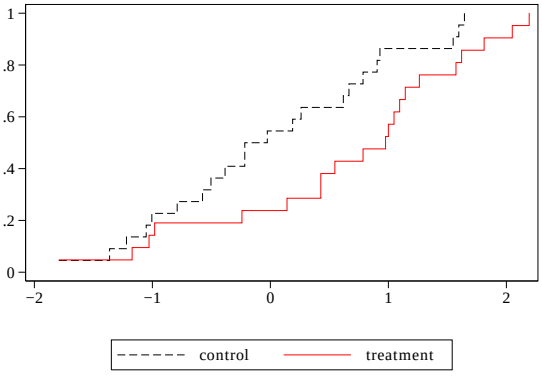
(b) CAT Age 7, Females
$p = .018$



(c) CAT Age 8, Males
$p = .253$



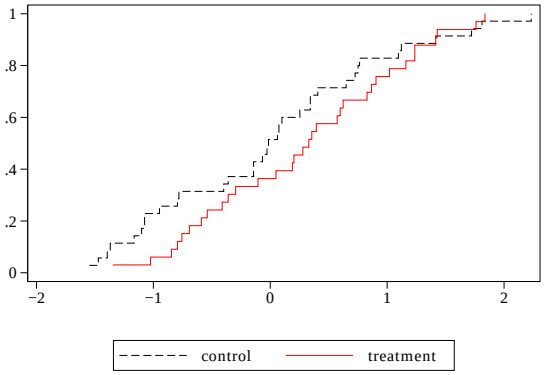
(d) CAT Age 8, Females
$p = .032$



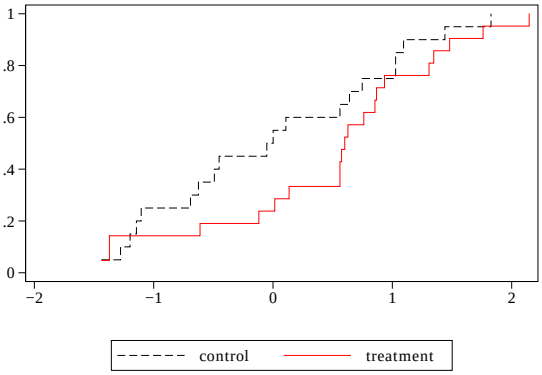
Notes: "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure B.4: Continued Empirical CDFs of the CAT Measures

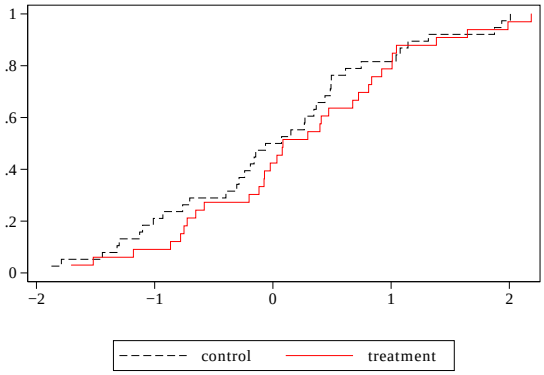
(e) CAT Age 9, Males
$p = .103$



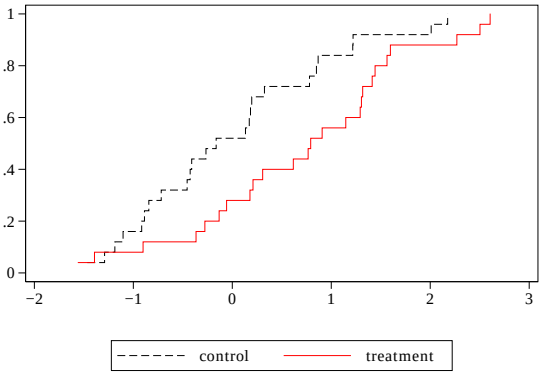
(f) CAT Age 9, Females
$p = .075$



(g) CAT Index, Males[a]
$p = .089$

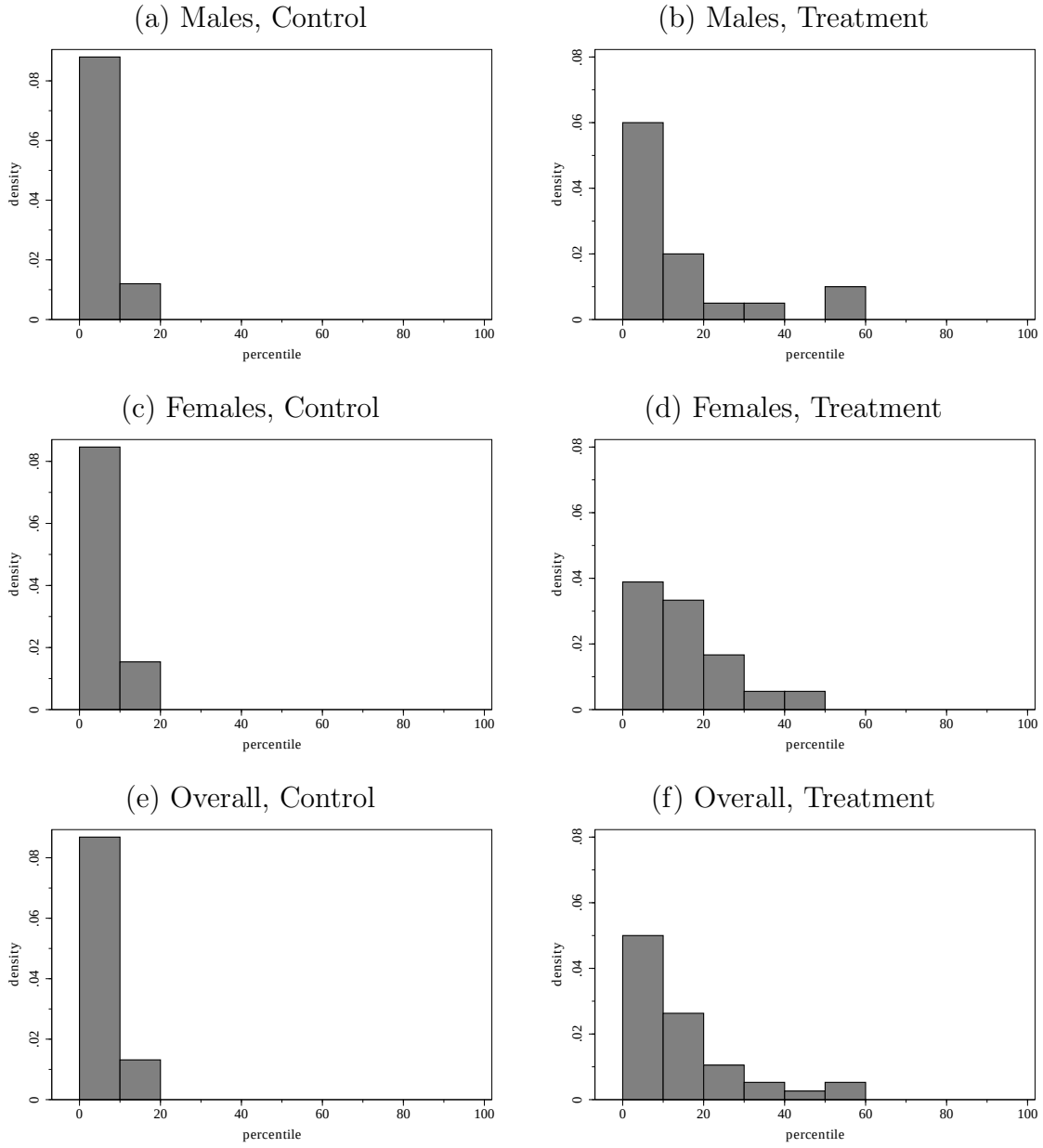


(h) CAT Index, Females[a]
$p = .023$



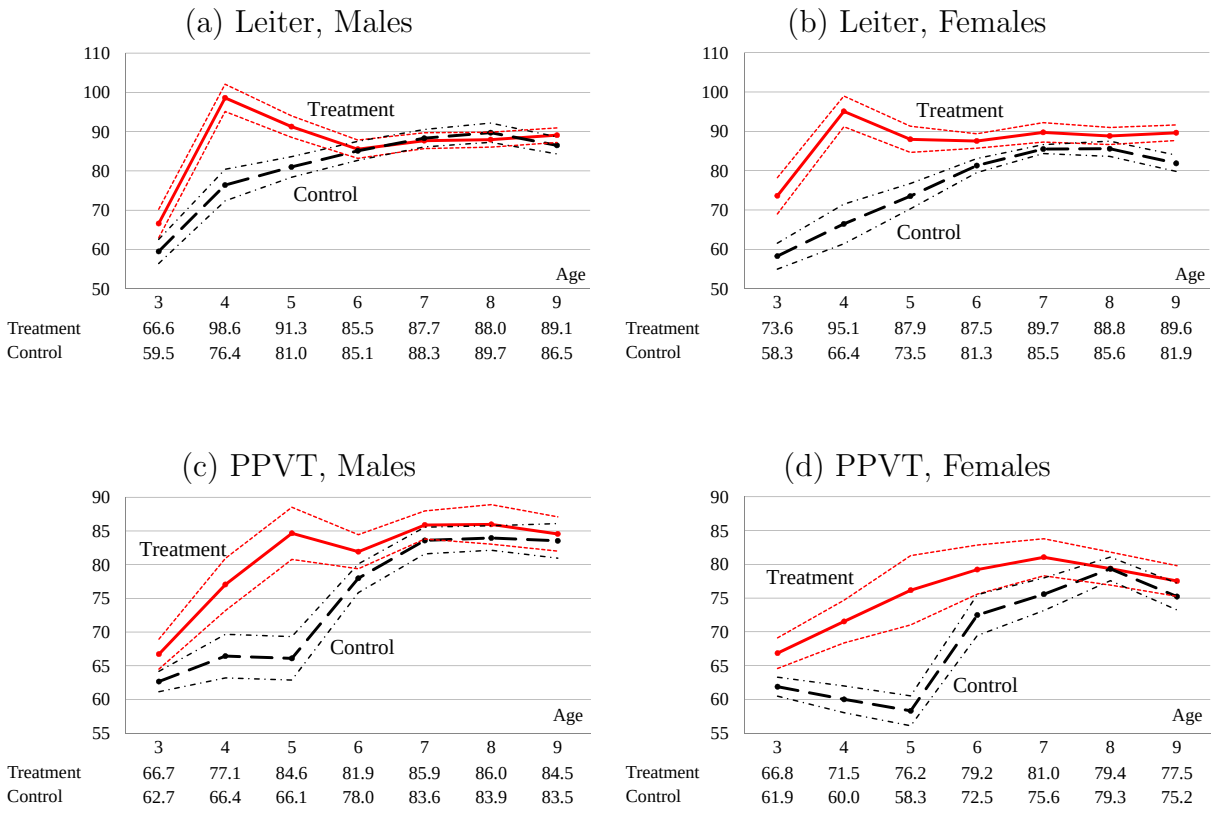
Notes: [a]The index is an average over ages 7, 8, and 9. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure B.5: Histograms of the CAT Total Score, Age 14



Notes: CAT is the California Achievement Test. Histograms show CAT scores measured in percentiles of general population scores. The one-sided p-values for difference in means are 0.016, 0.002, and 0.000 for samples of males, females, and pooled genders respectively.

Figure B.6: IQ Test Scores by Gender and Treatment Status[a]



[a]**Notes:** Leiter International Performance Scale (Leiter) and Peabody Picture Vocabulary Test (PPVT) scores are shown for the Perry sample. Bold lines represent mean IQs. Fine lines represent standard errors for the corresponding means (one standard error above and below). Numbers below each chart are treatment and control mean test scores.

C Pupil Behavior Inventory

The Pupil Behavior Inventory (PBI) was developed by Vinter et al. (1966) to measure behavioral and attitudinal factors that affect academic success. Weikart, Bond and McNeil (1978) analyze the effect of the Perry program on PBI scales. The measurement instrument consists of 34 items corresponding to five scales. The five PBI scales are as follows (with the number of proxying items shown in parentheses): "Academic Motivation" (9), "Classroom Conduct" (12), "Socio-Emotional State" (5), "Teacher Dependence" (2), and "Personal Behavior" (6) (see Table C.1 for the list of items grouped in the five scales).

PBI data were collected at ages 6, 7, 8, and 9. Teachers were given a list of behaviors and were asked to report the frequency with which each student performed each behavior: very frequently, frequently, sometimes, infrequently, or very infrequently. Unlike the YRS discussed below (see Appendix D), teachers were not explicitly asked to compare each student with his/her peers in the same class, and thus teachers likely compared each student against all students they had ever come into contact with. The answers were converted to a numerical scale (1–5), with higher numbers corresponding to more socially desirable behaviors such as more academic motivation or less stealing.

Table C.2 shows the polychoric longitudinal correlations among PBI items across ages 6–9. For each item, we estimate correlations between ages 6 and 7, 7 and 8, as well as 8 and 9. Individual correlations are statistically significant with a few rare exceptions. The joint test for the hypothesis that all three correlations are zero between subsequent years is rejected at the 5% level for all PBI items except PB1–21.[10,11]

Figures C.1–C.5 show empirical CDFs for all PBI items and for the indices based on

[10]The longitudinal structure of the Perry experiment allows us to obtain within-sample information necessary to impute missing data on measures. Many students who were not evaluated at a particular age were evaluated at ages close to the missing one. Assuming the stability of these measures over the period between ages 7 and 9, average scores for each person over non-missing items at ages 7, 8, and 9 were formed and used in analysis. By averaging, we not only augment the sample, but we also reduce the noisiness of the measures.

[11]Even though the correlation between ages 6 and 7 and the joint test are not statistically significant, the correlation is strongly statistically significant between ages 7, 8, and 9. We use only ages 7, 8, and 9 for estimation.

the five original PBI scales. We also report p-values for the difference in means between the treatment and control groups. While many treatment effects on items are statistically significant for females, far fewer are statistically significant for males. As a rule, statistically significant items for males are related either to Personal Behavior or to Classroom Conduct scales.

Table C.1: PBI Scales Description

Personal Behavior		Academic Motivation	
Absences or truancies	(C)	Shows initiatives	(C/E)
Inappropriate personal appearance	(C)	Alert and interested in school work	(O/C)
Lying or cheating	(C)	Learning retained well	(O)
Steals	(C)	Completes assignments	(C)
Swears or uses obscene words	(C)	Motivated toward academic performance	(O/C)
Poor personal hygiene	(C)	Positive concern for own education	(O/C)
		Hesitant to try, or gives up easily	(C)
Classroom Conduct		Uninterested in subject matter	(O)
Blames others for troubles	(C/N)	Shows positive leadership	(E)
Resistant to teachers	(C/A)		
Attempts to manipulate adults	(C/A)	Socio-Emotional State	
Influences others toward troublemaking	(E/A)	Appears depressed	(N)
Impulsive	(E/C)	Withdrawn and uncommunicative	(N)
Requires continuous supervision	(C)	Friendly and well-received by other pupils	(E)
Aggressive toward peers	(A)	Appears generally happy	(E)
Disobedient	(C)	Isolated, few or no friends	(E)
Easily led into trouble	(A/C)		
Resentful of criticism or discipline	(N)	Teacher Dependence	
Disrupts classroom procedures	(C/A)	Seeks constant reassurance	(N)
Teases or provokes students	(C/A)	Posessive of teacher	(N)

Notes: The table shows items that define five original PBI scales (Vinter et al., 1966). PBI scales are classified into five categories: Personal Behavior, Classroom Conduct, Academic Motivation, Socio-Emotional State and Teacher Dependency. In psychology, the most accepted theory on the classification of human personality is the Big Five Traits of Personality inventory. This theory classifies traits into five broad categories: Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A) and Neuroticism (N). PBI precedes the theory of the Big Five Traits of Personality, and thus PBI categories do not match the Big Five traits perfectly. We thanks Angela Lee Duckworth for classifying each PBI measure in terms of the Big Five traits of Personality. The classification is presented in parenthesis.

Table C.2: Polychoric Longitudinal Correlations Among PBI Items Across Ages

Item	Description	Statistic	6-7	7-8	8-9	joint test
		corr	0.341	0.402	0.274	0.259
PBI 1	Shows Initiative	*p* -value	***0.002***	***0.001***	***0.080***	***0.000***
		N	97	78	52	
		corr	0.451	0.392	0.363	0.252
PBI 2	Blames others for troubles	*p* -value	***0.000***	***0.001***	***0.012***	***0.000***
		N	98	80	55	
		corr	0.169	0.241	0.326	0.192
PBI 3	Resistant to teachers	*p* -value	0.141	***0.052***	***0.026***	***0.001***
		N	97	79	55	
		corr	0.383	0.441	0.455	0.180
PBI 4	Alert and interested in school work	*p* -value	***0.000***	***0.000***	***0.002***	***0.002***
		N	98	80	54	
		corr	0.264	0.145	0.440	0.205
PBI 5	Attempts to manipulate adults	*p* -value	***0.018***	0.256	***0.003***	***0.001***
		N	96	77	53	
		corr	0.038	0.506	0.565	0.203
PBI 6	Appears depressed	*p* -value	0.736	***0.000***	***0.000***	***0.001***
		N	98	80	55	
		corr	0.494	0.638	0.489	0.214
PBI 7	Learning retained well	*p* -value	***0.000***	***0.000***	***0.001***	***0.000***
		N	98	80	55	
		corr	0.523	0.432	0.649	0.180
PBI 8	Absences or truancies	*p* -value	***0.000***	***0.001***	***0.000***	***0.003***
		N	98	80	55	
		corr	0.204	0.400	0.506	0.153
PBI 9	Withdrawn and uncommunicative	*p* -value	***0.065***	***0.001***	***0.000***	***0.010***
		N	98	80	55	
		corr	0.401	0.439	0.397	0.207
PBI 10	Completes assignments	*p* -value	***0.000***	***0.000***	***0.007***	***0.000***
		N	98	80	55	
		corr	0.479	0.339	0.271	0.225
PBI 11	Influences others toward troublemaking	*p* -value	***0.000***	***0.006***	***0.064***	***0.000***
		N	98	80	55	
		corr	0.201	0.373	0.374	0.238
PBI 12	Inappropriate personal appearance	*p* -value	***0.093***	***0.005***	***0.023***	***0.000***
		N	98	80	55	
		corr	0.308	0.111	0.368	0.212
PBI 13	Seeks constant reassurance	*p* -value	***0.006***	0.378	***0.010***	***0.000***
		N	98	80	55	
		corr	0.455	0.339	0.549	0.191
PBI 14	Motivated toward academic performance	*p* -value	***0.000***	***0.005***	***0.000***	***0.001***
		N	98	79	53	
		corr	0.078	0.305	0.353	0.182
PBI 15	Impulsive	*p* -value	0.486	***0.016***	***0.017***	***0.002***
		N	97	79	54	
		corr	0.280	0.237	0.369	0.141
PBI 16	Lying or cheating	*p* -value	***0.019***	***0.074***	***0.014***	***0.021***
		N	88	70	50	
		corr	0.346	0.519	0.285	0.128
PBI 17	Positive concern for own education	*p* -value	***0.003***	***0.000***	***0.059***	***0.028***
		N	89	73	54	

Notes: Polychoric correlations across PBI measures at subsequent ages (6 and 7, 7 and 8, 8 and 9), p-values, and sample sizes are shown. p-values are for the likelihood ratio test of no correlation. p-values that are below 10% are in bold.

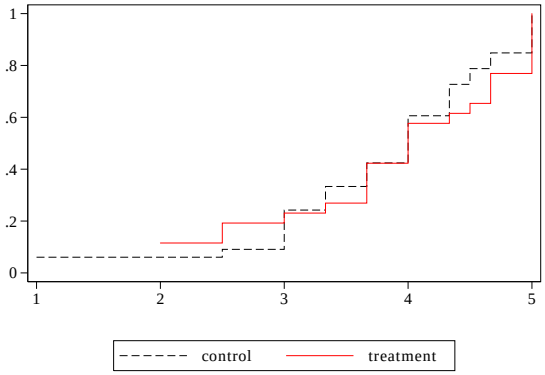
Table C.2: Continued Polychoric Longitudinal Correlations Among PBI Items Across Ages

Item	Description	Statistic	6-7	7-8	8-9	joint test
		corr	0.230	0.286	0.419	0.161
PBI 18	Requires continuous supervision	*p* -value	***0.035***	***0.019***	***0.003***	***0.006***
		N	98	79	54	
		corr	0.245	-0.011	0.242	0.128
PBI 19	Aggressive toward peers	*p* -value	***0.025***	0.931	0.115	***0.029***
		N	98	78	52	
		corr	0.513	0.500	0.577	0.264
PBI 20	Disobedient	*p* -value	***0.000***	***0.000***	***0.000***	***0.000***
		N	98	80	54	
		corr	0.041	0.480	0.454	0.055
PBI 21	Steals	*p* -value	0.777	***0.001***	***0.014***	0.411
		N	81	65	45	
		corr	0.100	0.233	0.282	0.132
PBI 22	Friendly and well-received by other pupils	*p* -value	0.386	***0.063***	***0.054***	***0.026***
		N	98	80	55	
		corr	0.419	0.424	0.498	0.247
PBI 23	Easily led into trouble	*p* -value	***0.000***	***0.000***	***0.000***	***0.000***
		N	98	80	54	
		corr	0.142	0.392	0.462	0.209
PBI 24	Resentful of criticism or discipline	*p* -value	0.208	***0.001***	***0.001***	***0.000***
		N	98	80	55	
		corr	0.220	0.546	0.451	0.278
PBI 25	Hesitant to try, or gives up easily	*p* -value	***0.042***	***0.000***	***0.002***	***0.000***
		N	98	80	55	
		corr	0.338	0.565	0.360	0.228
PBI 26	Uninterested in subject matter	*p* -value	***0.002***	***0.000***	***0.013***	***0.000***
		N	98	80	55	
		corr	0.408	0.503	0.508	0.220
PBI 27	Disrupts classroom procedures	*p* -value	***0.000***	***0.000***	***0.000***	***0.000***
		N	98	80	55	
		corr	0.339	0.522	0.486	0.144
PBI 28	Swears or uses obscene words	*p* -value	***0.007***	***0.000***	***0.001***	***0.025***
		N	84	68	49	
		corr	0.185	0.435	0.434	0.243
PBI 29	Appears generally happy	*p* -value	0.101	***0.000***	***0.005***	***0.000***
		N	98	80	55	
		corr	0.333	0.487	0.464	0.248
PBI 30	Poor personal hygiene	*p* -value	***0.004***	***0.000***	***0.002***	***0.000***
		N	98	80	55	
		corr	0.113	0.299	0.306	0.227
PBI 31	Possessive of teacher	*p* -value	0.328	***0.014***	***0.048***	***0.000***
		N	97	80	53	
		corr	0.497	0.223	0.378	0.527
PBI 32	Teases or provokes students	*p* -value	***0.000***	***0.087***	***0.010***	***0.000***
		N	98	70	52	
		corr	0.084	0.344	0.509	0.508
PBI 33	Isolated, few or no friends	*p* -value	0.468	***0.008***	***0.000***	***0.000***
		N	96	70	52	
		corr	0.444	0.502	0.347	0.498
PBI 34	Shows positive leadership	*p* -value	***0.000***	***0.000***	***0.026***	***0.000***
		N	97	69	52	

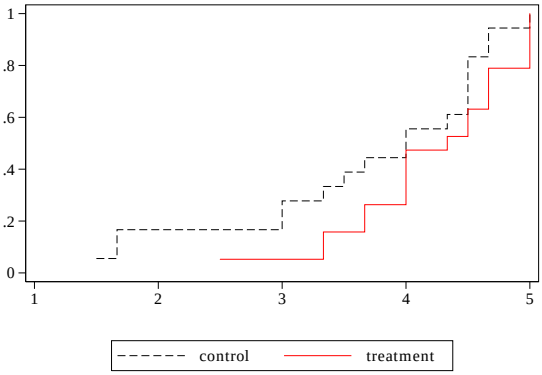
Notes: Polychoric correlations across PBI measures at subsequent ages (6 and 7, 7 and 8, 8 and 9), p-values, and sample sizes are shown. p-values are for the likelihood ratio test of no correlation. p-values that are below 10% are in bold.

Figure C.1: Empirical CDFs of the PBI Personal Behavior Items

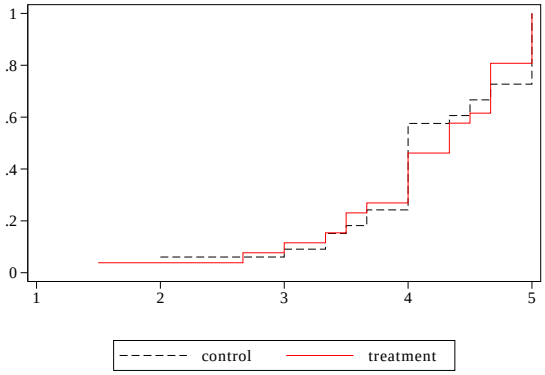
(a) Absences or Truancies, Males
$p = .374$



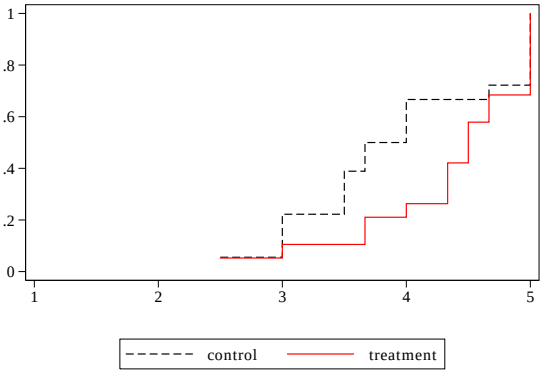
(b) Absences or Truancies, Females
$p = .042$



(c) Inappropriate Personal Appearance, Males
$p = .520$



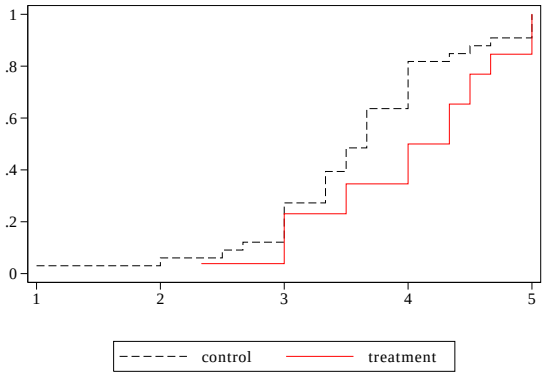
(d) Inappropriate Personal Appearance, Females
$p = .059$



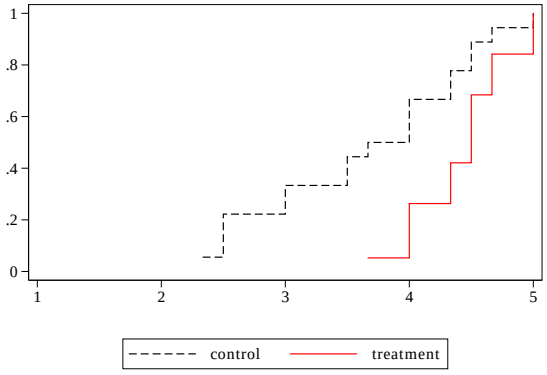
Notes: Each PBI item in these charts is an average over non-missing observations at ages 7, 8, and 9. Each item takes values from 1 to 5, with higher numbers corresponding to more socially-desirable behaviors like more learning or less stealing. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure C.1: Continued Empirical CDFs of the PBI Personal Behavior Items

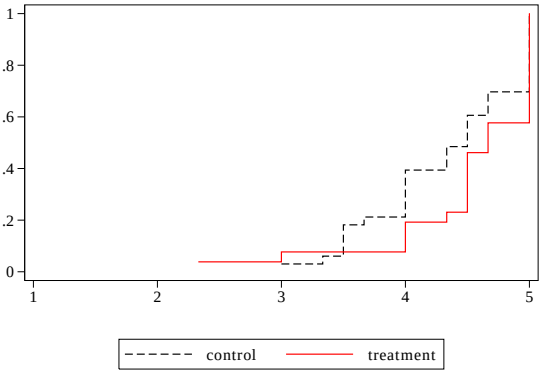
(e) Lying or Cheating, Males
$p = .024$



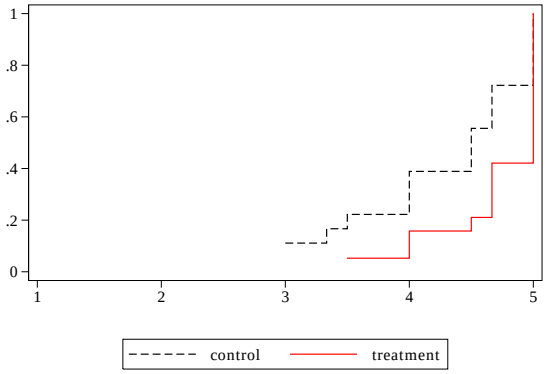
(f) Lying or Cheating, Females
$p = .000$



(g) Steals, Males
$p = .122$



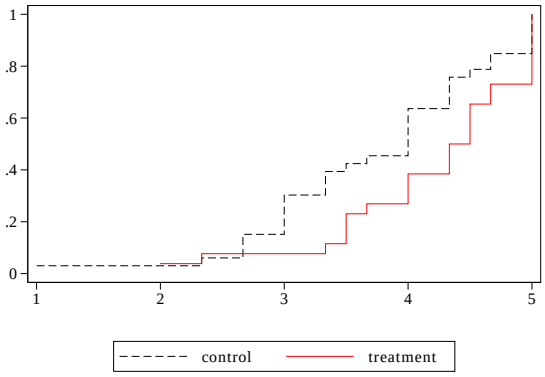
(h) Steals, Females
$p = .016$



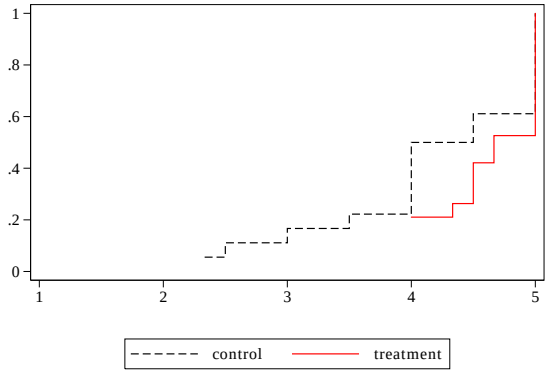
Notes: Each PBI item in these charts is an average over non-missing observations at ages 7, 8, and 9. Each item takes values from 1 to 5, with higher numbers corresponding to more socially-desirable behaviors like more learning or less stealing. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure C.1: Continued Empirical CDFs of the PBI Personal Behavior Items

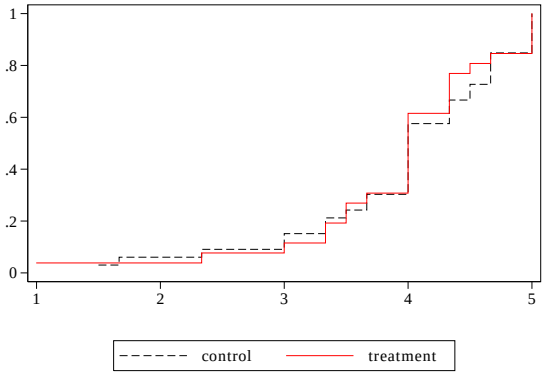
(i) Swears or Uses Obscene Words, Males
$p = .028$



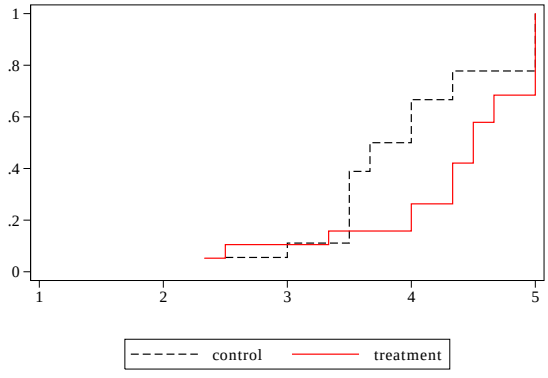
(j) Swears or Uses Obscene Words, Females
$p = .025$



(k) Poor Personal Hygienes, Males
$p = .551$



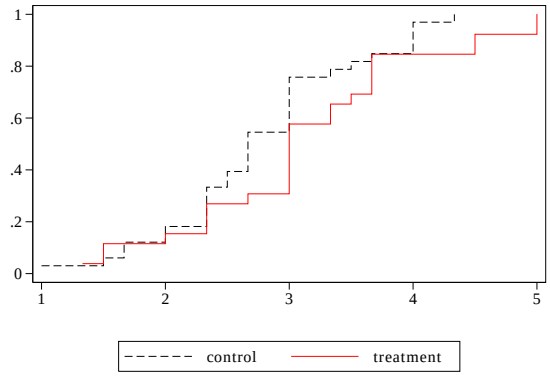
(l) Poor Personal Hygienes, Females
$p = .074$



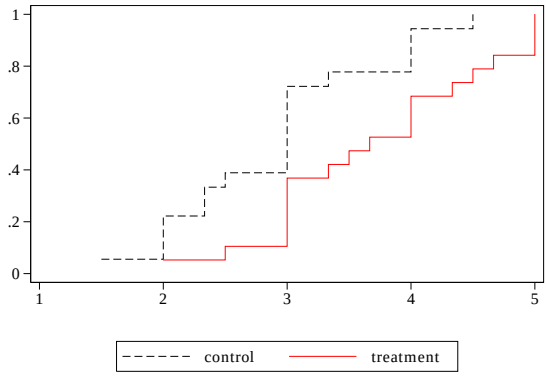
Notes: Each PBI item in these charts is an average over non-missing observations at ages 7, 8, and 9. Each item takes values from 1 to 5, with higher numbers corresponding to more socially-desirable behaviors like more learning or less stealing. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure C.2: Empirical CDFs of the PBI Classroom Conduct Items

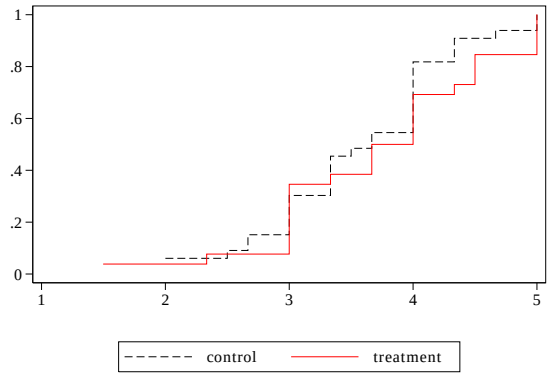
(a) Blames Others for Troubles, Males
$p = .078$



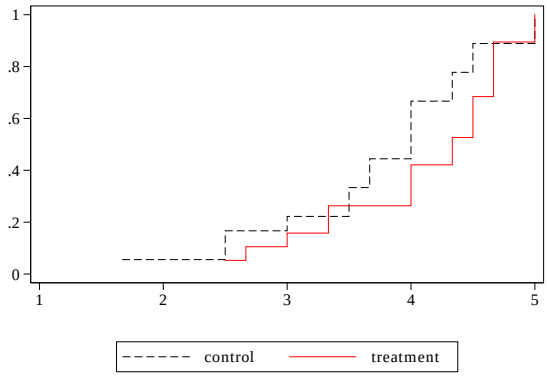
(b) Blames Others for Troubles, Females
$p = .004$



(c) Resistant to Teachers, Males
$p = .236$



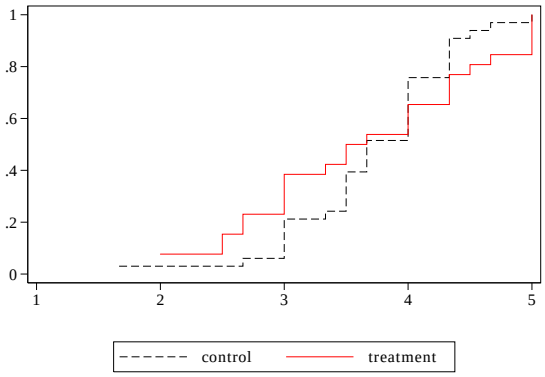
(d) Resistant to Teachers, Females
$p = .117$



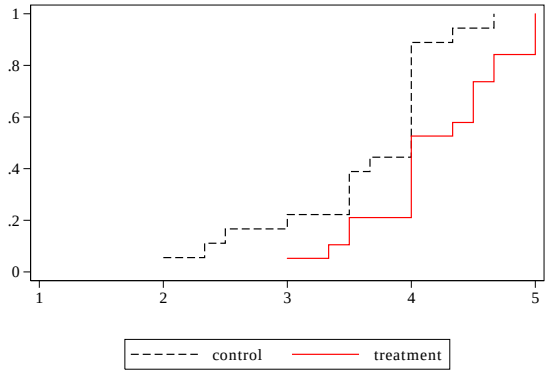
Notes: Each PBI item in these charts is an average over non-missing observations at ages 7, 8, and 9. Each item takes values from 1 to 5, with higher numbers corresponding to more socially-desirable behaviors like more learning or less stealing. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure C.2: Continued Empirical CDFs of the PBI Classroom Conduct Items

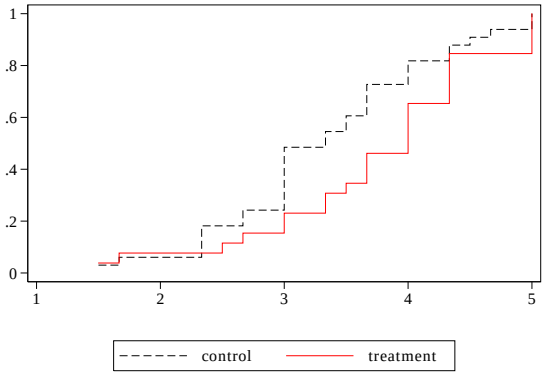
(e) Attempts to Manipulate Adults, Males
$p = .652$



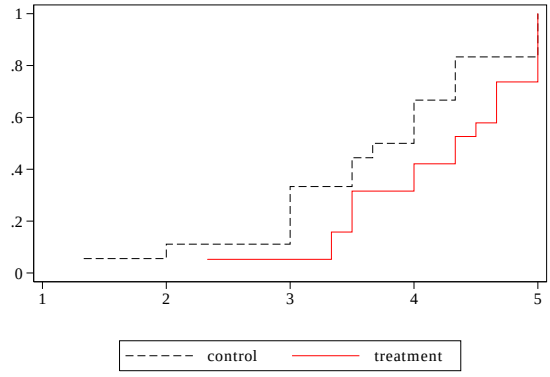
(f) Attempts to Manipulate Adults, Females
$p = .006$



(g) Influences Others Toward Trouble Making, Males
$p = .047$



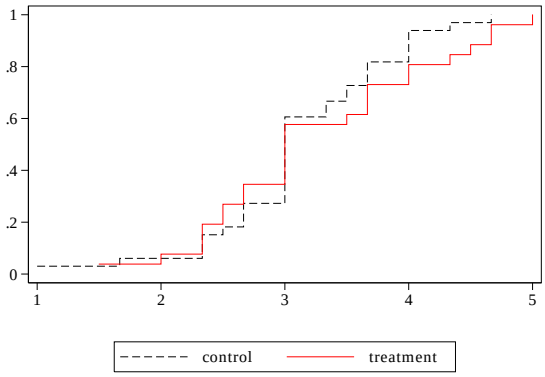
(h) Influences Others Toward Trouble Making, Females
$p = .040$



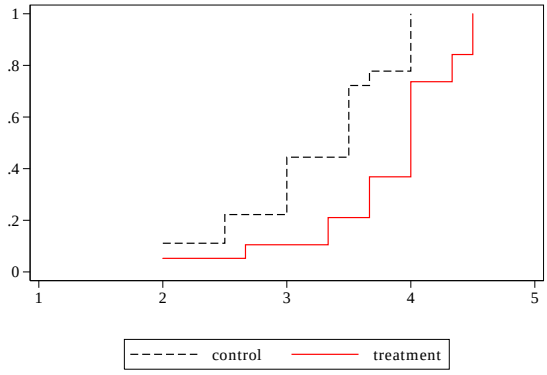
Notes: Each PBI item in these charts is an average over non-missing observations at ages 7, 8, and 9. Each item takes values from 1 to 5, with higher numbers corresponding to more socially-desirable behaviors like more learning or less stealing. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure C.2: Continued Empirical CDFs of the PBI Classroom Conduct Items

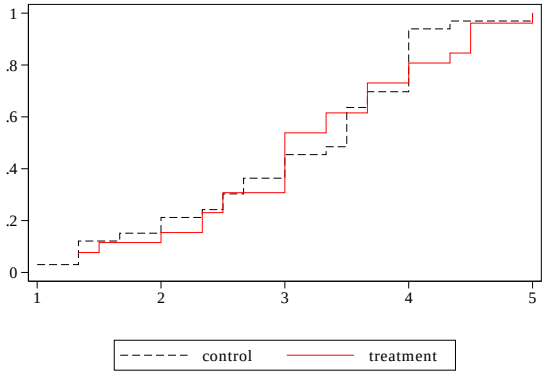
(i) Impulsive, Males
$p = .269$



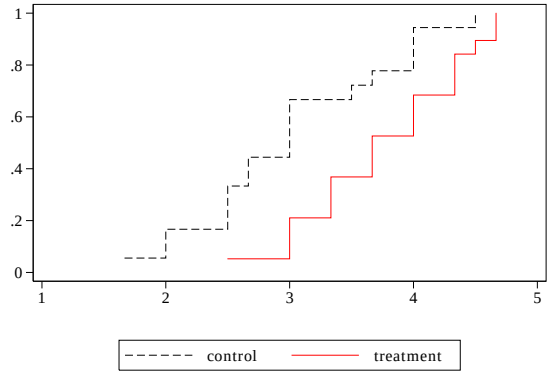
(j) Impulsive, Females
$p = .005$



(k) Requires Continuous Supervision, Males
$p = .381$



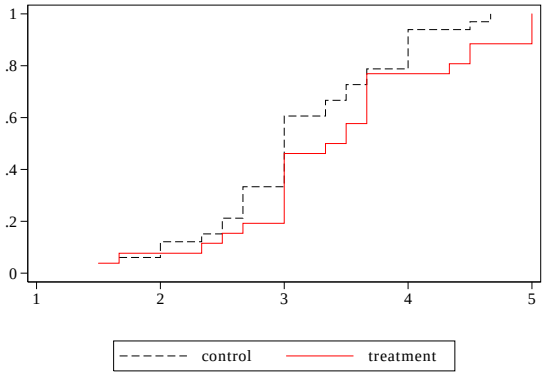
(l) Requires Continuous Supervision, Females
$p = .002$



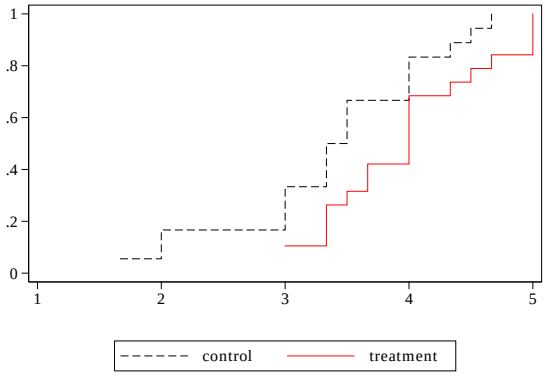
Notes: Each PBI item in these charts is an average over non-missing observations at ages 7, 8, and 9. Each item takes values from 1 to 5, with higher numbers corresponding to more socially-desirable behaviors like more learning or less stealing. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure C.2: Continued Empirical CDFs of the PBI Classroom Conduct Items

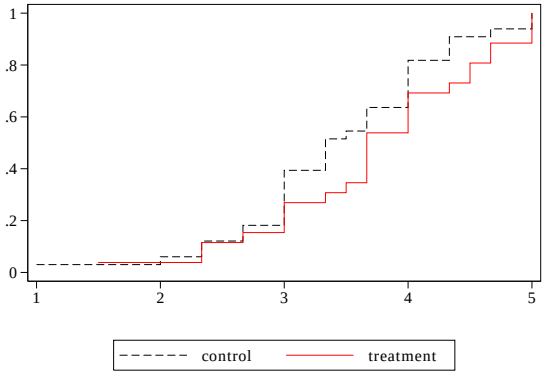
(m) Aggressive Toward Peers, Males
$p = .085$



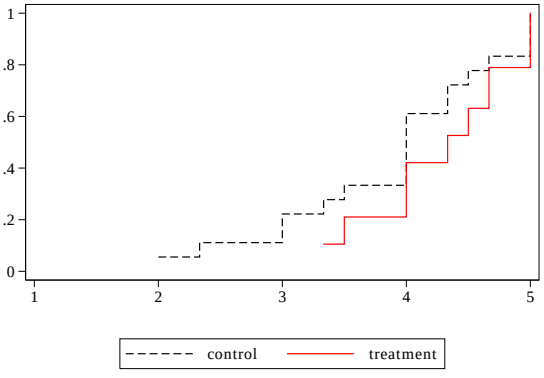
(n) Aggressive Toward Peers, Females
$p = .011$



(o) Disobedient, Males
$p = .116$



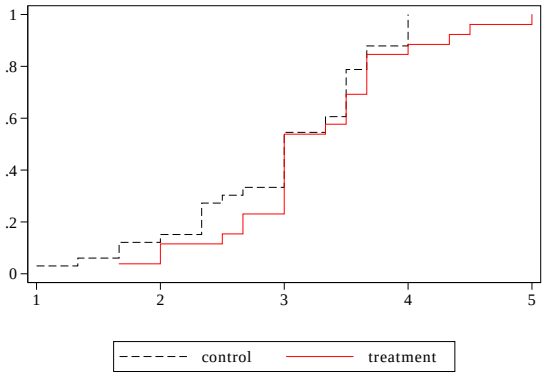
(p) Disobedient, Females
$p = .058$



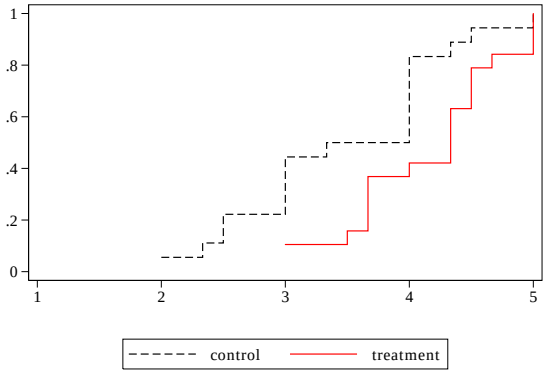
Notes: Each PBI item in these charts is an average over non-missing observations at ages 7, 8, and 9. Each item takes values from 1 to 5, with higher numbers corresponding to more socially-desirable behaviors like more learning or less stealing. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure C.2: Continued Empirical CDFs of the PBI Classroom Conduct Items

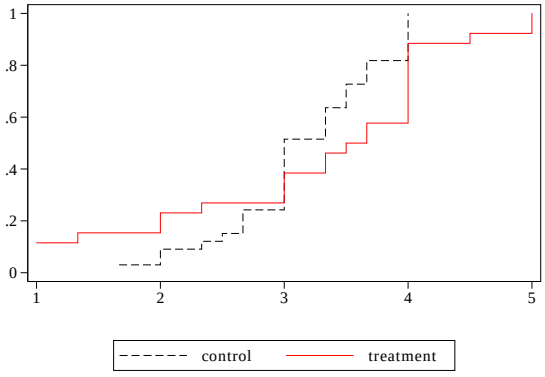
(q) Easily Led into Trouble, Males
$p = .106$



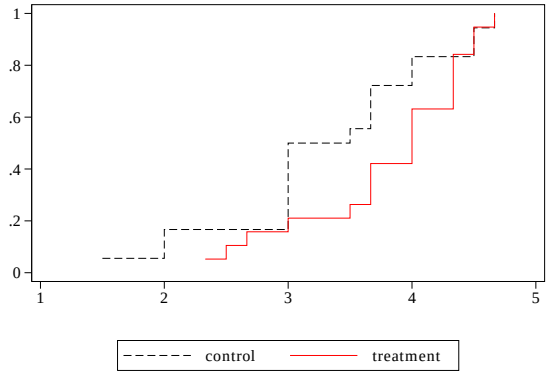
(r) Easily Led into Trouble, Females
$p = .005$



(s) Resentful of Criticism or Discipline, Males
$p = .408$



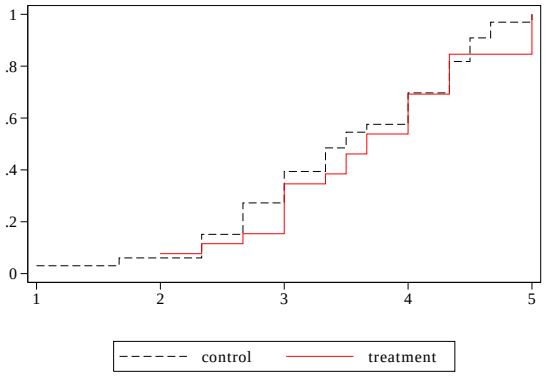
(t) Resentful of Criticism or Discipline, Females
$p = .039$



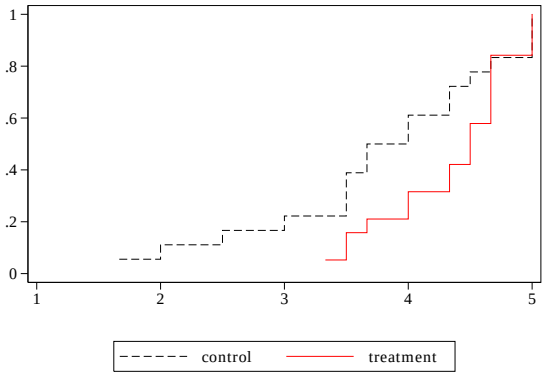
Notes: Each PBI item in these charts is an average over non-missing observations at ages 7, 8, and 9. Each item takes values from 1 to 5, with higher numbers corresponding to more socially-desirable behaviors like more learning or less stealing. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure C.2: Continued Empirical CDFs of the PBI Classroom Conduct Items

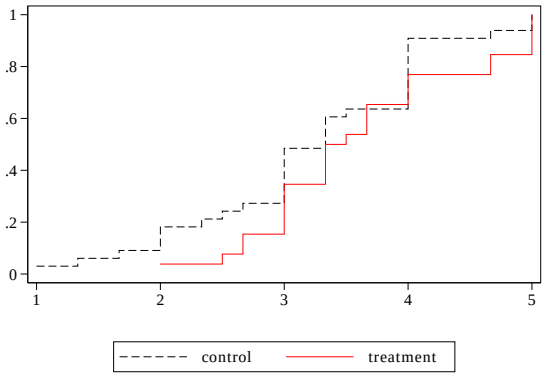
(u) Disrupts Classroom Procedures, Males
$p = .215$



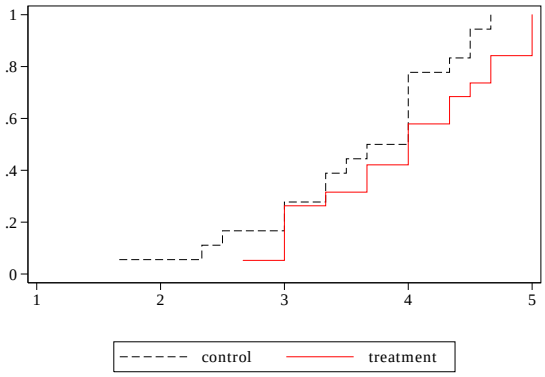
(v) Disrupts Classroom Procedures, Females
$p = .017$



(w) Teases or Provokes Students, Males
$p = .052$



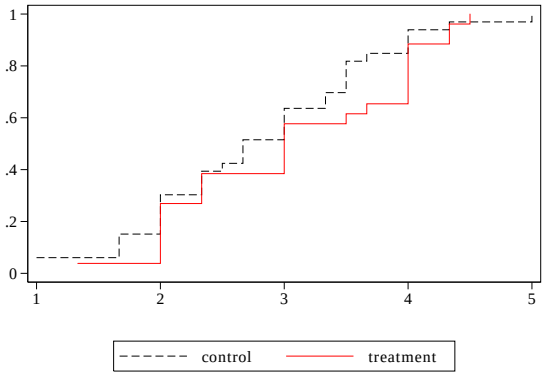
(x) Teases or Provokes Students, Females
$p = .087$



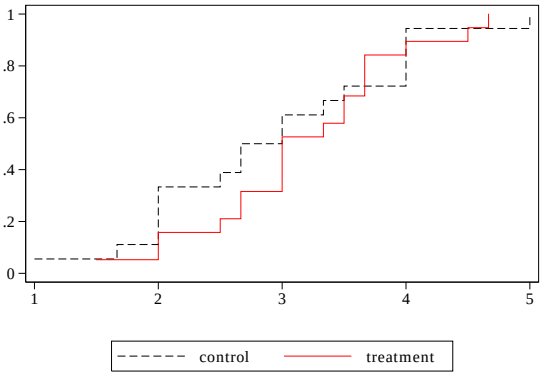
Notes: Each PBI item in these charts is an average over non-missing observations at ages 7, 8, and 9. Each item takes values from 1 to 5, with higher numbers corresponding to more socially-desirable behaviors like more learning or less stealing. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure C.3: Empirical CDFs of the PBI Academic Motivation Items

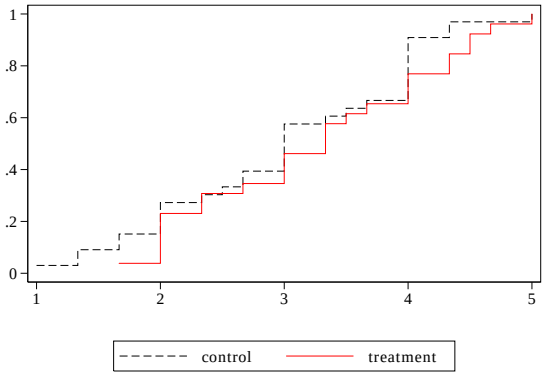
(a) Shows Initiative, Males
$p = .141$



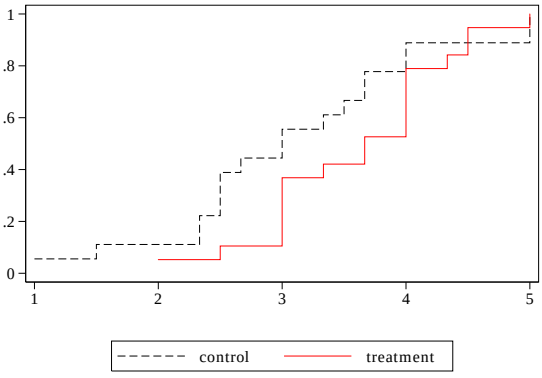
(b) Shows Initiative, Females
$p = .219$



(c) Alert and Interested in Schoolwork, Males
$p = .187$



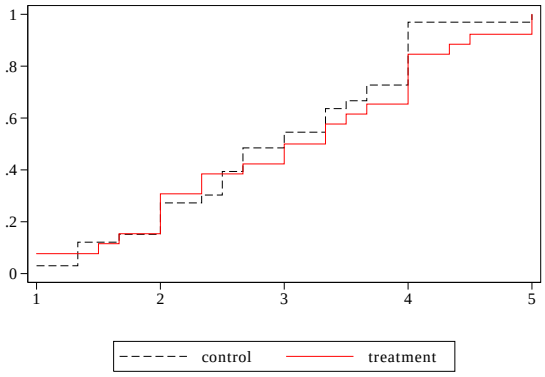
(d) Alert and Interested in Schoolwork, Females
$p = .047$



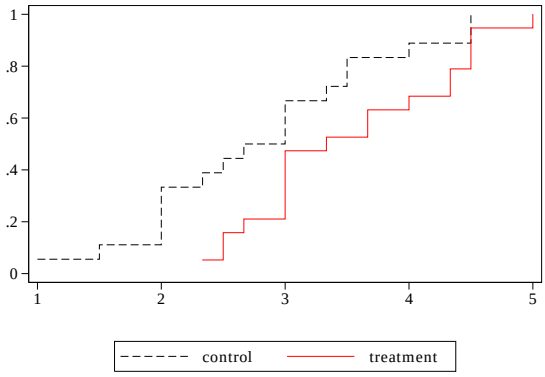
Notes: Each PBI item in these charts is an average over non-missing observations at ages 7, 8, and 9. Each item takes values from 1 to 5, with higher numbers corresponding to more socially-desirable behaviors like more learning or less stealing. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure C.3: Continued Empirical CDFs of the PBI Academic Motivation Items

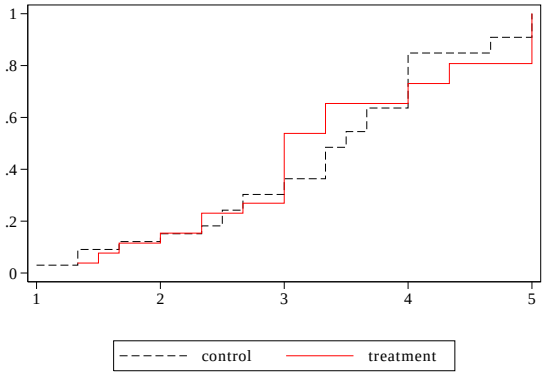
(e) Learning Retained Well, Males
$p = .331$



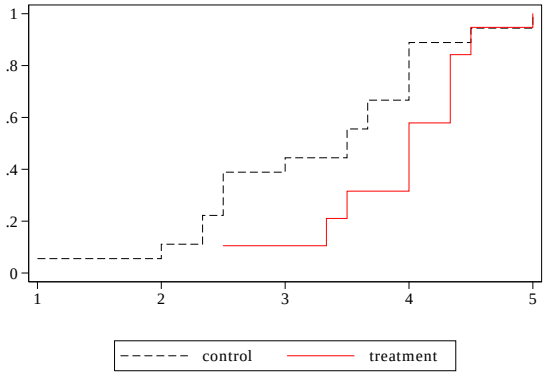
(f) Learning Retained Well, Females
$p = .010$



(g) Completes Assignments, Males
$p = .495$



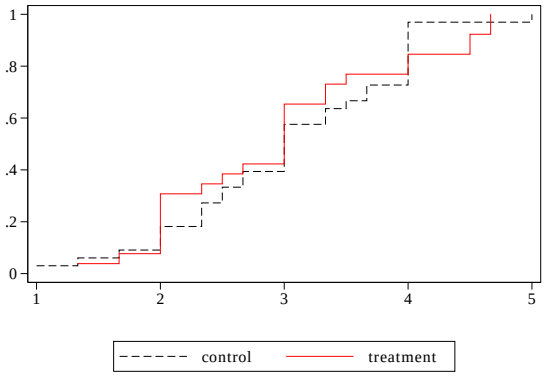
(h) Completes Assignments, Females
$p = .009$



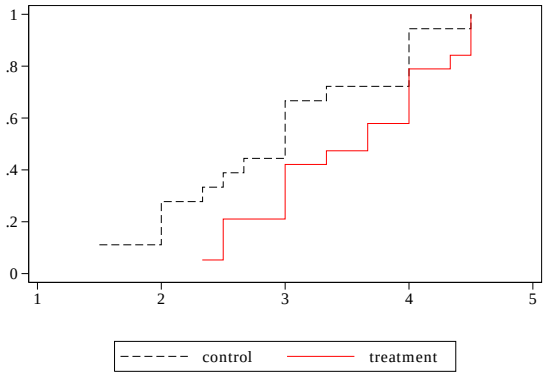
Notes: Each PBI item in these charts is an average over non-missing observations at ages 7, 8, and 9. Each item takes values from 1 to 5, with higher numbers corresponding to more socially-desirable behaviors like more learning or less stealing. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure C.3: Continued Empirical CDFs of the PBI Academic Motivation Items

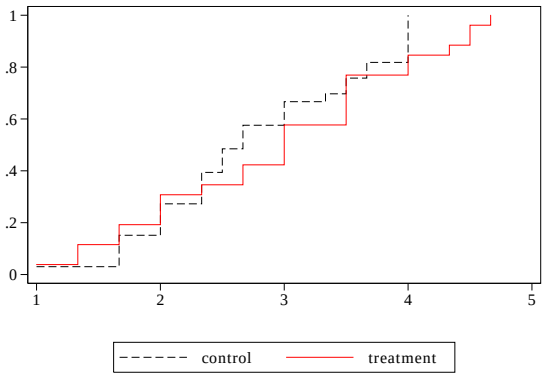
(i) Motivated Toward Academic Performance, Males
$p = .601$



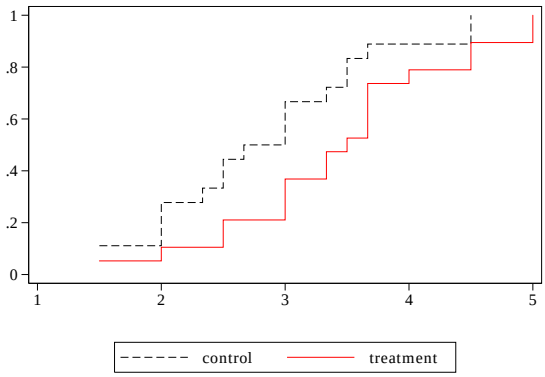
(j) Motivated Toward Academic Performance, Females
$p = .021$



(k) Positive Concern for Own Education, Males
$p = .255$



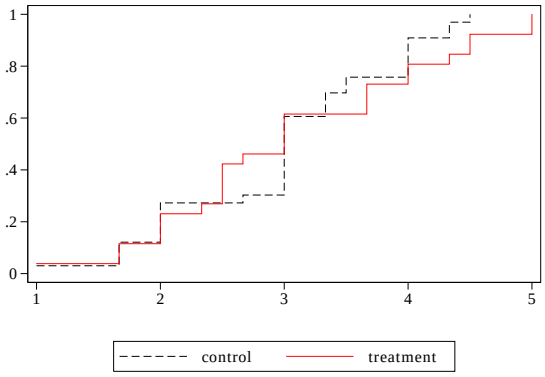
(l) Positive Concern for Own Education, Females
$p = .026$



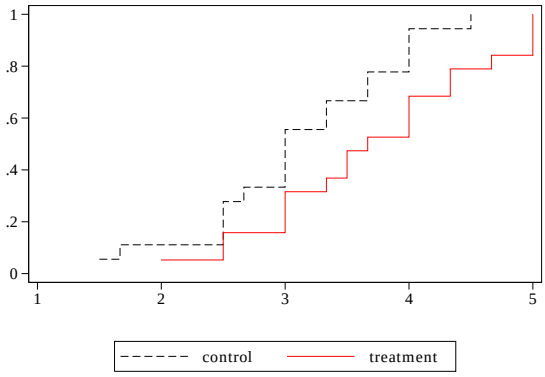
Notes: Each PBI item in these charts is an average over non-missing observations at ages 7, 8, and 9. Each item takes values from 1 to 5, with higher numbers corresponding to more socially-desirable behaviors like more learning or less stealing. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure C.3: Continued Empirical CDFs of the PBI Academic Motivation Items

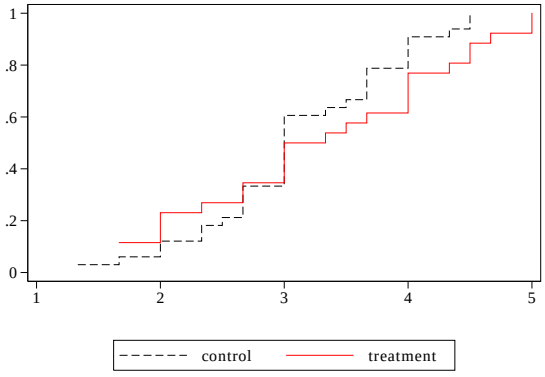
(m) Hesitant to Try, or Gives Up Easily, Males
$p = .395$



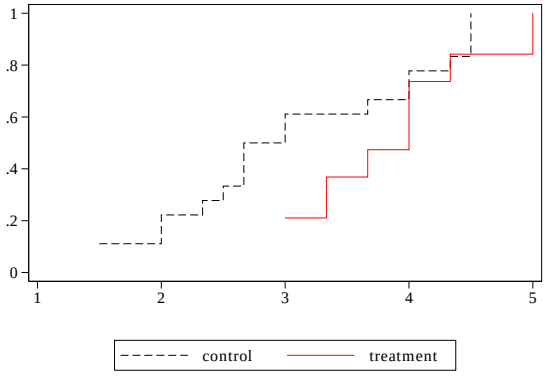
(n) Hesitant to Try, or Gives Up Easily, Females
$p = .020$



(o) Uninterested in Subject Matter, Males
$p = .251$



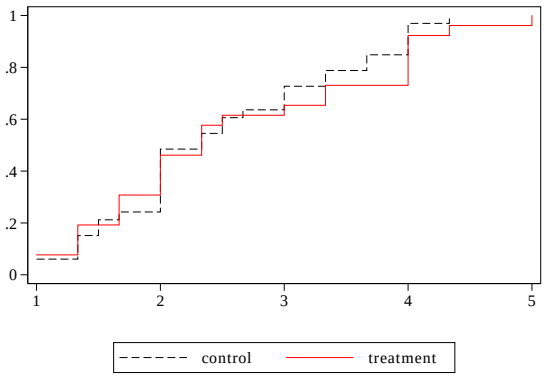
(p) Uninterested in Subject Matter, Females
$p = .006$



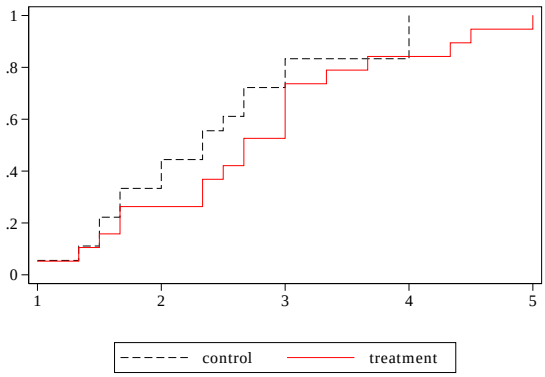
Notes: Each PBI item in these charts is an average over non-missing observations at ages 7, 8, and 9. Each item takes values from 1 to 5, with higher numbers corresponding to more socially-desirable behaviors like more learning or less stealing. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure C.3: Continued Empirical CDFs of the PBI Academic Motivation Items

(q) Shows Positive Leadership, Males
$p = .359$



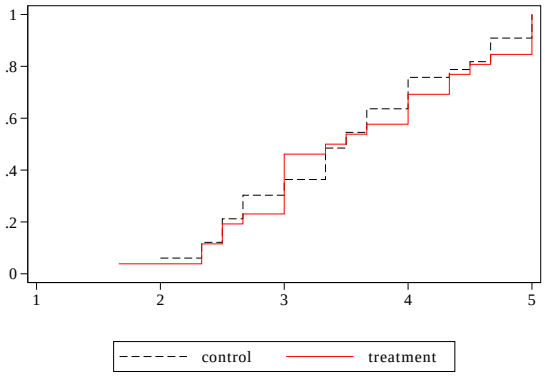
(r) Shows Positive Leadership, Females
$p = .141$



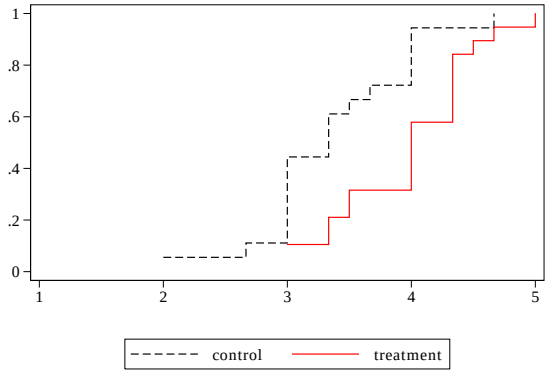
Notes: Each PBI item in these charts is an average over non-missing observations at ages 7, 8, and 9. Each item takes values from 1 to 5, with higher numbers corresponding to more socially-desirable behaviors like more learning or less stealing. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure C.4: Empirical CDFs of the PBI Socio-Emotional State Items

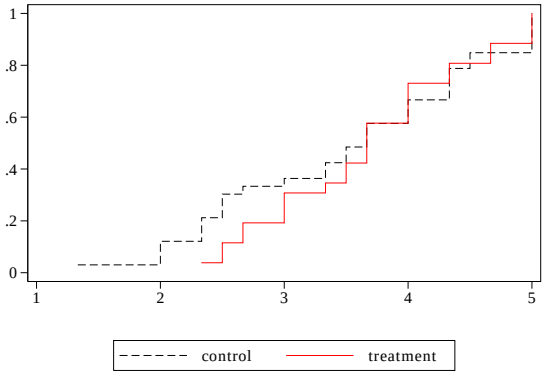
(a) Appears Depressed, Males
$p = .410$



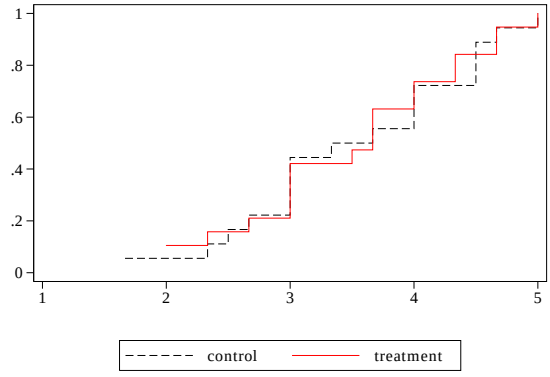
(b) Appears Depressed, Females
$p = .002$



(c) Withdrawn and Uncommunicative, Males
$p = .240$



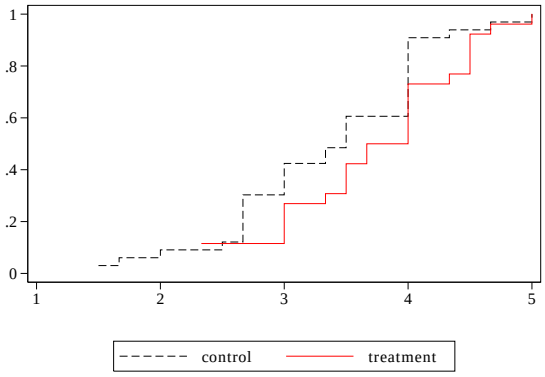
(d) Withdrawn and Uncommunicative, Females
$p = .524$



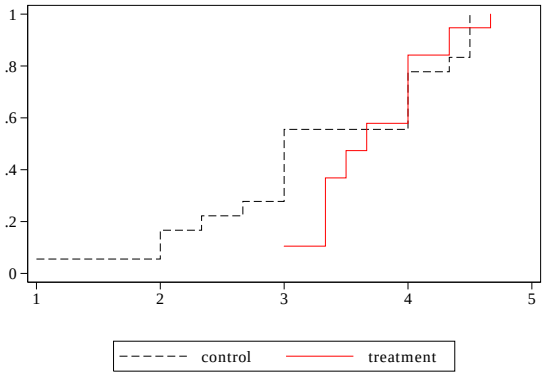
Notes: Each PBI item in these charts is an average over non-missing observations at ages 7, 8, and 9. Each item takes values from 1 to 5, with higher numbers corresponding to more socially-desirable behaviors like more learning or less stealing. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure C.4: Continued Empirical CDFs of the PBI Socio-Emotional State Items

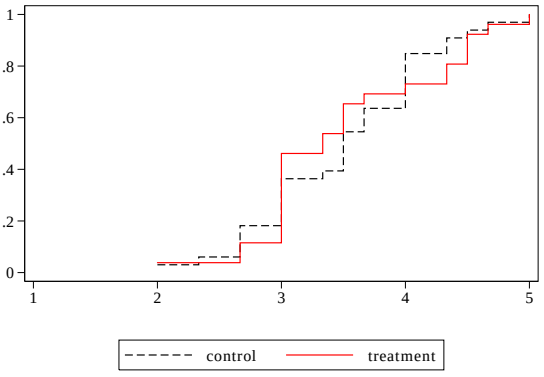
(e) Friendly and Well-Received by by Other Pupils, Males
$p = .046$



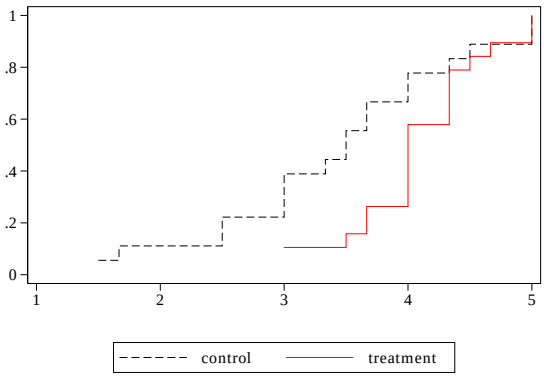
(f) Friendly and Well-Received by Other Pupils, Females
$p = .052$



(g) Appears Generally Happy, Males
$p = .511$



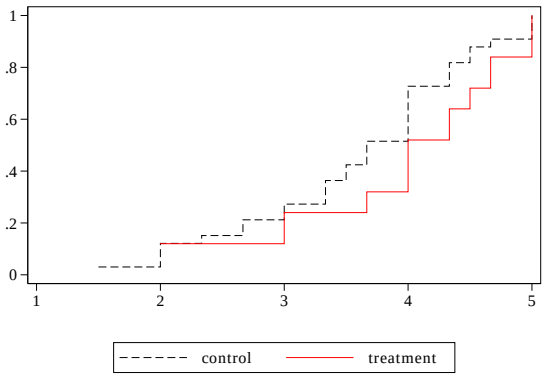
(h) Appears Generally Happy, Females
$p = .010$



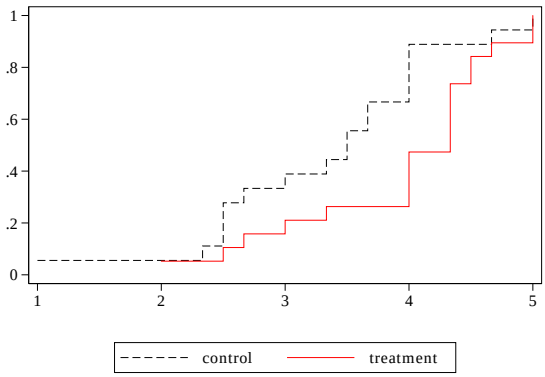
Notes: Each PBI item in these charts is an average over non-missing observations at ages 7, 8, and 9. Each item takes values from 1 to 5, with higher numbers corresponding to more socially-desirable behaviors like more learning or less stealing. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure C.4: Continued Empirical CDFs of the PBI Socio-Emotional State Items

(i) Isolated, Few or no Friends, Males
$p = .093$



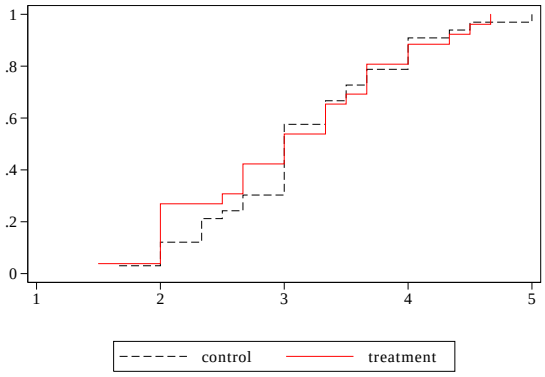
(j) Isolated, Few or no Friends, Females
$p = .023$



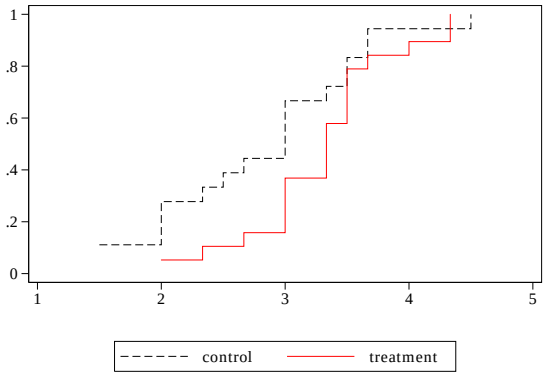
Notes: Each PBI item in these charts is an average over non-missing observations at ages 7, 8, and 9. Each item takes values from 1 to 5, with higher numbers corresponding to more socially-desirable behaviors like more learning or less stealing. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure C.5: Empirical CDFs of the PBI Teacher Dependence Items

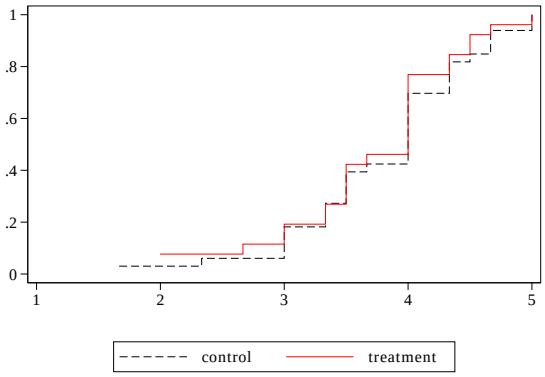
(a) Seeks Constant Reassurance, Males
$p = .681$



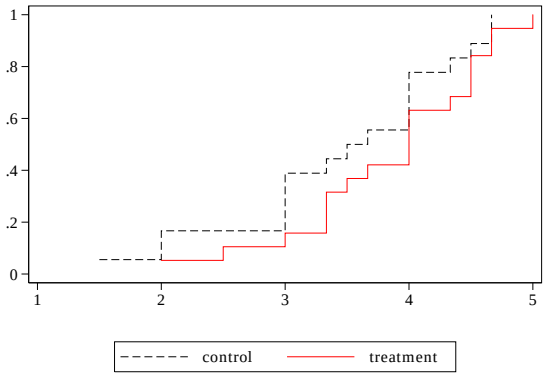
(b) Seeks Constant Reassurance, Females
$p = .023$



(c) Possessive of Teacher, Males
$p = .692$

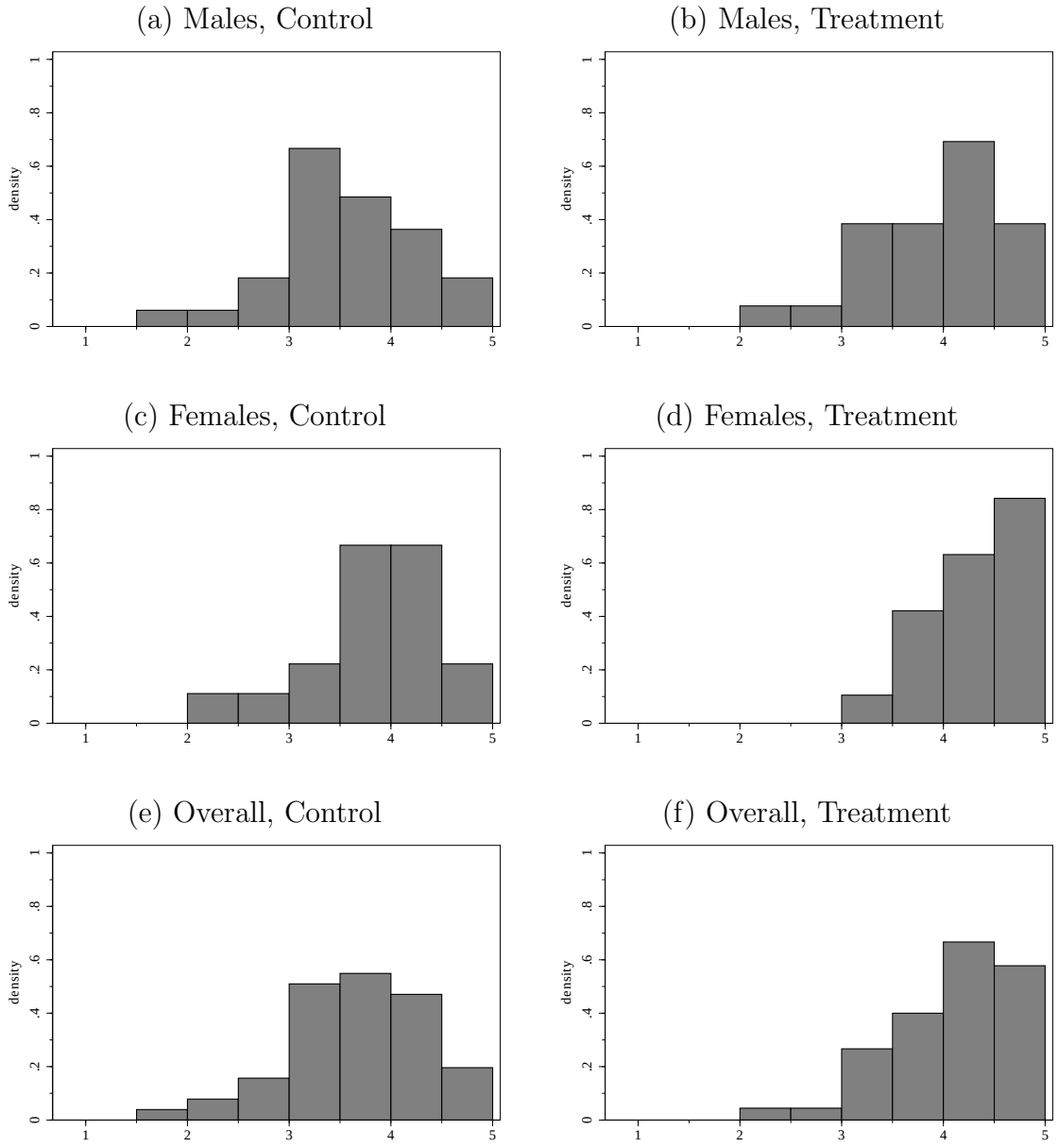


(d) Possessive of Teacher, Females
$p = .095$



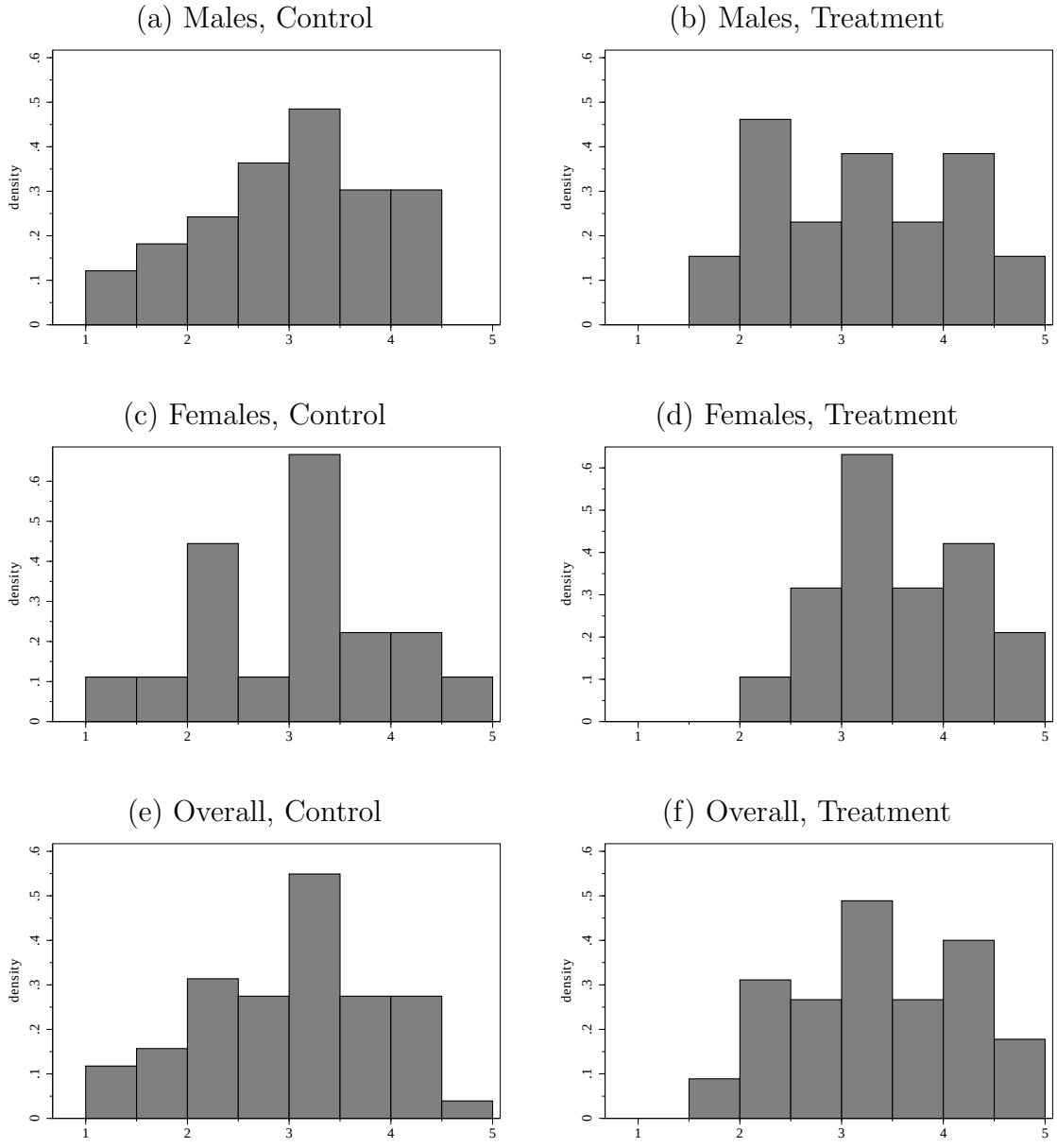
Notes: Each PBI item in these charts is an average over non-missing observations at ages 7, 8, and 9. Each item takes values from 1 to 5, with higher numbers corresponding to more socially-desirable behaviors like more learning or less stealing. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure C.6: Histograms of Externalizing Behavior Index



Notes: The Externalizing Behavior index is an unweighted average of seven measures: "disrupts classroom procedures," "swears or uses obscene words," "steals," "lying or cheating," "influences others toward troublemaking," "aggressive toward peers," and "teases or provokes students". Higher numbers correspond to more socially-desirable behaviors. The one-sided p-values for difference in means are 0.031, 0.006, and 0.001 for samples of males, females, and pooled genders respectively.

Figure C.7: Histograms of Academic Motivation Index



Notes: The Academic Motivation index is an unweighted average of three measures: "shows initiative," "alert and interested in school work," and "hesitant to try, or gives up easily,". Higher numbers correspond to more socially-desirable behaviors. The one-sided p-values for difference in means are 0.211, 0.053, and 0.043 for samples of males, females, and pooled genders respectively.

D Ypsilanti Rating Scale

The Ypsilanti Rating Scale[12] (YRS) was developed by the Perry project staff (Weikart, Bond and McNeil, 1978) as an additional measures of personality and school readiness. The 9 YRS items relevant to personality skills are listed in Table D.1. These items define four scales (with the number of proxying items shown in parentheses): "Academic Potential" (3), "Social Development" (3), "Verbal Skills" (1), and "Emotional Adjustment" (2).

Data for the YRS were collected at ages 6, 7, 8, and 9. Teachers were instructed to compare each child to other students in a specified small group. Teachers ranked the students on a scale from 1–7, with higher scores corresponding to more socially-desirable behaviors or skills.[13]

Table D.2 shows the polychoric longitudinal correlations between ages 6 and 7, 7 and 8, as well as 8 and 9 for the nine YRS items. Individual correlations are generally statistically significant with some exceptions. The joint test for the hypothesis that all three correlations between subsequent years is always rejected at the 5% level.

Figures D.1–D.4 show empirical CDFs for individual items and indexes of the YRS scale. For males, all treatment effects on measures are not statistically significant. For females, some treatment effects related to Academic Potential, Social Development, and Emotional Adjustment are statistically significant.

[12] We considered using the YRS scales to estimate the model, but following the analysis reported in Web Appendix H, we only use items from the PBI scales as measures of personality skills.

[13] The longitudinal structure of the Perry experiment allows us to obtain within-sample information necessary to solve the problem of missing data on measures. Students who were not evaluated at a particular age were often evaluated at ages close to the age (or ages) with the missing data. Assuming the stability of these measures over the period between ages 7 and 9, average scores for each person over non-missing observations at ages 7, 8, and 9 were formed and used in the analysis. By averaging, we not only augment the sample, but also reduce the noisiness of the measures.

Table D.1: YRS Scales Description

Academic Potential		Social Development	
Degree of imagination and creativity shown	(O)	Social relationship with classmates	(A/E/C)
Level of academic readiness	(C/A/O/IQ)	Social relationship with teachers	(A/C)
Prediction of future academic success	(C/A/O/IQ)	Level of curiosity shown	(O)
Verbal Skill		Emotional Adjustment	
Level of verbal communication	(IQ)	Level of emotional adjustment	(N)
		Degree of trust of total environment	(A/N)

Notes: The table shows items that define five original YRS scales. YRS scales are classified into four categories: Academic Potential, Verbal Skill, Social Development and Emotional Adjustment. In psychology, the most accepted theory on the classification of human personality is given by the Big Five Traits of Personality inventory. This theory classifies traits into five broad categories: Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A) and Neuroticism (N). YRS precedes the theory of the Big Five Traits of Personality, and thus YRS categories do not match the Big Five traits perfectly. We thank Angela Duckworth for classifying each YRS measure in terms of the Big Five traits of Personality and IQ. The classification is presented in parenthesis.

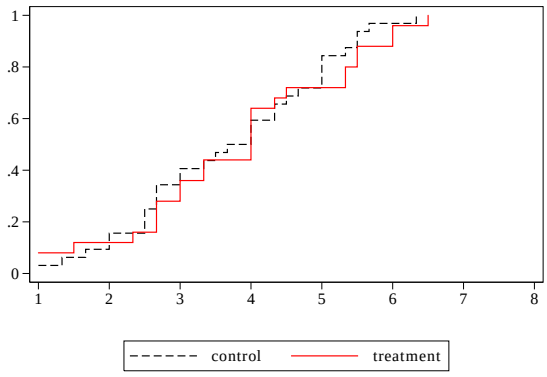
Table D.2: Polychoric Longitudinal Correlations Among YRS Items Across Subsequent Ages

Description	Statistic	6-7	7-8	8-9	Joint Test
Social relationship with classmates	*corr*	0.263	0.601	0.297	0.386
	P-value	***0.021***	***0.000***	***0.018***	***0.000***
	N	82	67	68	
Social relationship with teachers	*corr*	0.225	0.254	0.201	0.237
	P-value	***0.051***	***0.040***	0.110	***0.001***
	N	82	67	68	
Level of verbal communication	*corr*	0.429	0.459	0.462	0.443
	P-value	***0.000***	***0.000***	***0.000***	***0.000***
	N	81	66	68	
Degree of imagination and creativity shown	*corr*	0.364	0.474	0.197	0.356
	P-value	***0.001***	***0.000***	0.125	***0.000***
	N	82	67	68	
Level of academic readiness	*corr*	0.563	0.478	0.559	0.535
	P-value	***0.000***	***0.000***	***0.000***	***0.000***
	N	82	67	68	
Level of curiosity shown	*corr*	0.280	0.593	0.196	0.341
	P-value	***0.015***	***0.000***	0.120	***0.000***
	N	81	66	68	
Level of emotional adjustment	*corr*	0.226	0.369	0.479	0.353
	P-value	***0.050***	***0.003***	***0.000***	***0.000***
	N	82	67	68	
Prediction of future academic success	*corr*	0.538	0.601	0.601	0.587
	P-value	***0.000***	***0.000***	***0.000***	***0.000***
	N	82	67	68	
Degree of trust of total environment	*corr*	0.118	0.281	0.225	0.161
	P-value	0.325	***0.025***	***0.072***	***0.023***
	N	81	67	68	

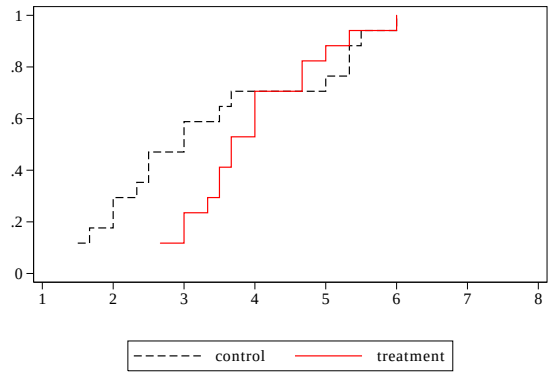
Notes: Polychoric correlations among PBI measures at subsequent ages (6 and 7, 7 and 8, 8 and 9), p-values, and sample sizes are shown. p-values are for the likelihood ratio test of no correlation. p-values that are below 10% are in bold.

Figure D.1: Empirical CDFs of the Academic Potential YRS Measures

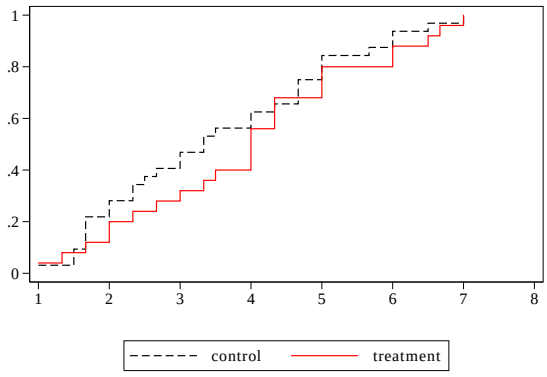
(a) Degree of Imagination and Creativity Shown, Males
$p = .448$



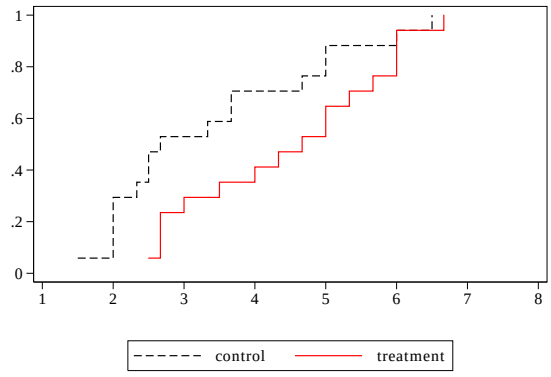
(b) Degree of Imagination and Creativity Shown, Females
$p = .125$



(c) Level of Academic Readiness, Males
$p = .357$



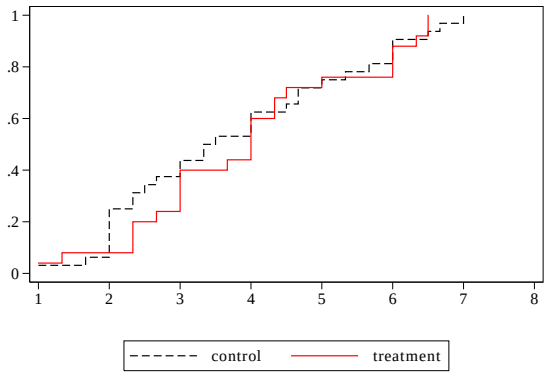
(d) Level of Academic Readiness, Females
$p = .183$



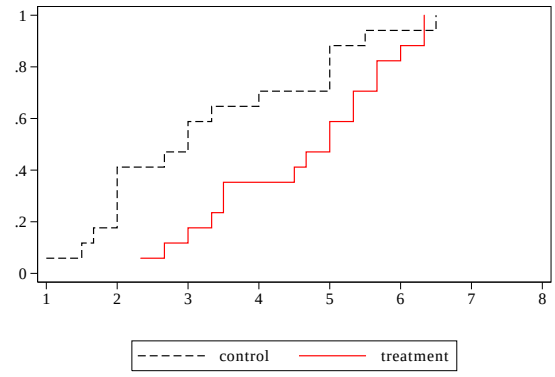
Notes: Each YRS item in these charts is an average over non-missing observations at ages 7, 8, and 9. Each item takes values from 1 to 7, with higher numbers corresponding to more socially-desirable behaviors like more learning or less stealing. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure D.1: Continued Empirical CDFs of the Academic Potential YRS Measures

(e) Prediction of Future Academic Success, Males
$p = .589$



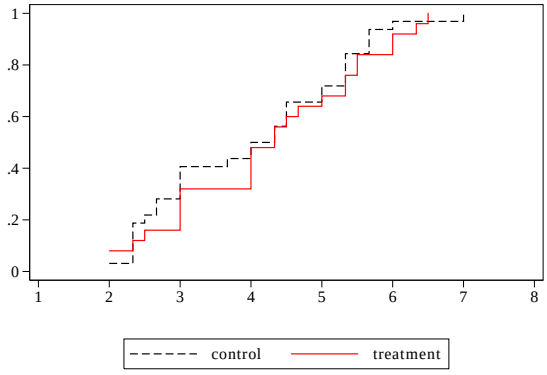
(f) Prediction of Future Academic Success, Females
$p = .062$



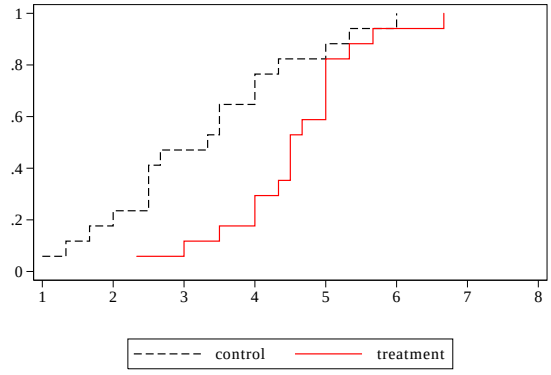
Notes: Each YRS item in these charts is an average over non-missing observations at ages 7, 8, and 9. Each item takes values from 1 to 7, with higher numbers corresponding to more socially-desirable behaviors like more learning or less stealing. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure D.2: Empirical CDFs of the Social Development YRS Measures

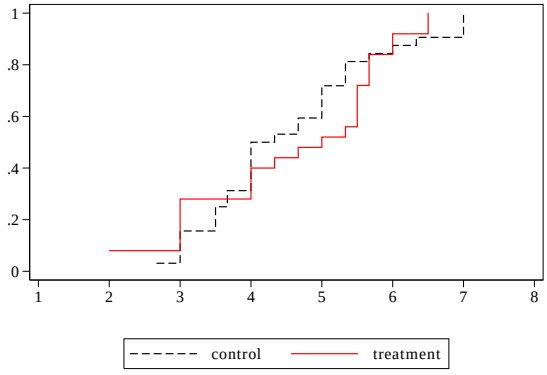
(a) Social Relationship with Classmates, Males
$p = .271$



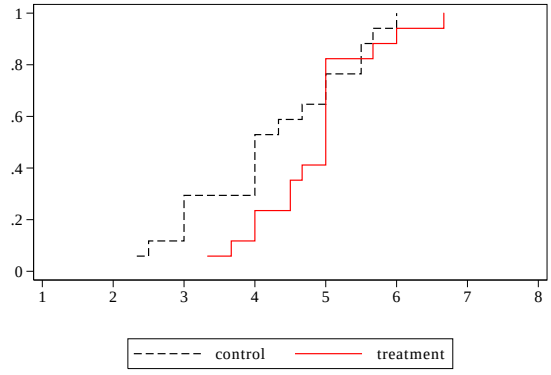
(b) Social Relationship with Classmates, Females
$p = .002$



(c) Social Relationship with Teachers, Males
$p = .458$



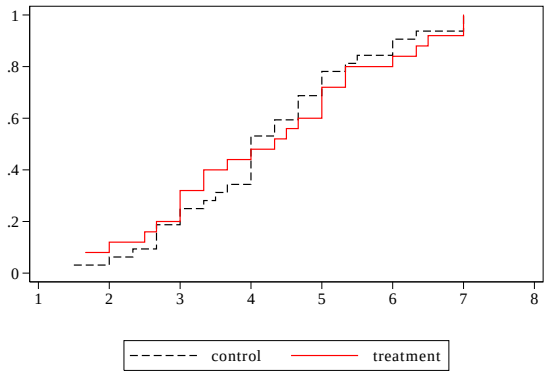
(d) Social Relationship with Teachers, Females
$p = .041$



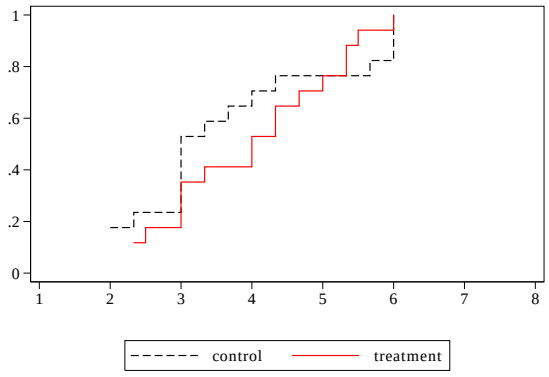
Notes: Each YRS item in these charts is an average over non-missing observations at ages 7, 8, and 9. Each item takes values from 1 to 7, with higher numbers corresponding to more socially-desirable behaviors like more learning or less stealing. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure D.2: Continued Empirical CDFs of the Social Development YRS Measures

(e) Level of Curiosity Shown, Males
$p = .491$



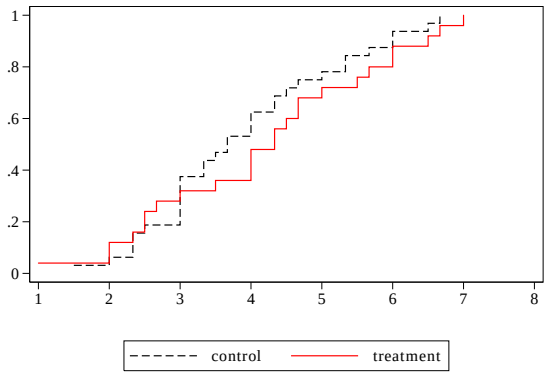
(f) Level of Curiosity Shown, Females
$p = .234$



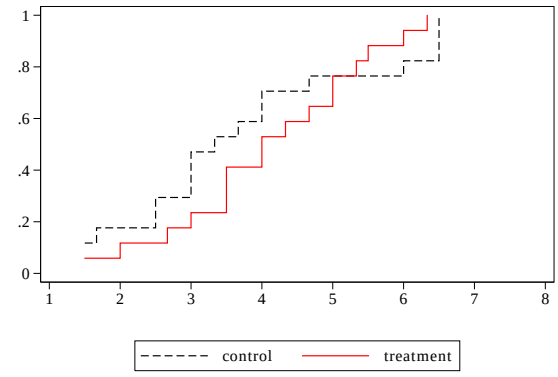
Notes: Each YRS item in these charts is an average over non-missing observations at ages 7, 8, and 9. Each item takes values from 1 to 7, with higher numbers corresponding to more socially-desirable behaviors like more learning or less stealing. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure D.3: Empirical CDFs of the Verbal Skills YRS Measures

(a) Level of Verbal Communication, Males
$p = .247$



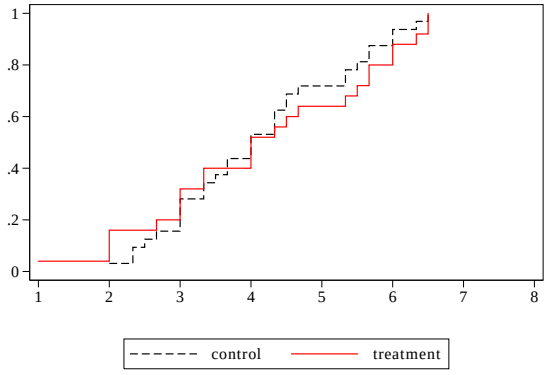
(b) Level of Verbal Communication, Females
$p = .257$



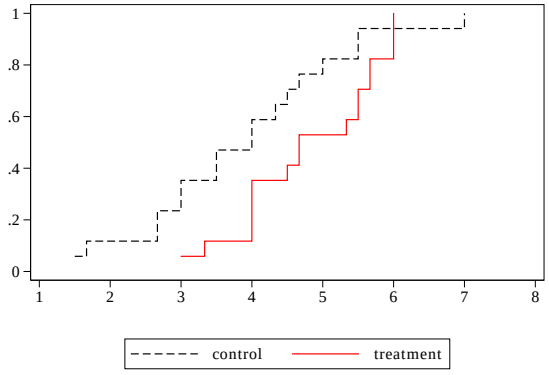
Notes: Each YRS item in these charts is an average over non-missing observations at ages 7, 8, and 9. Each item takes values from 1 to 7, with higher numbers corresponding to more socially-desirable behaviors like more learning or less stealing. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

Figure D.4: Empirical CDFs of the Emotional Adjustment YRS Measures

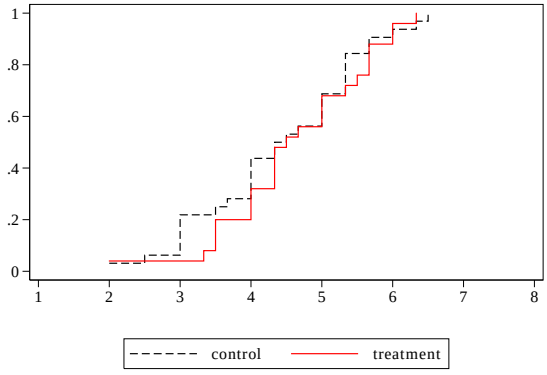
(a) Level of Emotional Adjustments, Males
$p = .461$



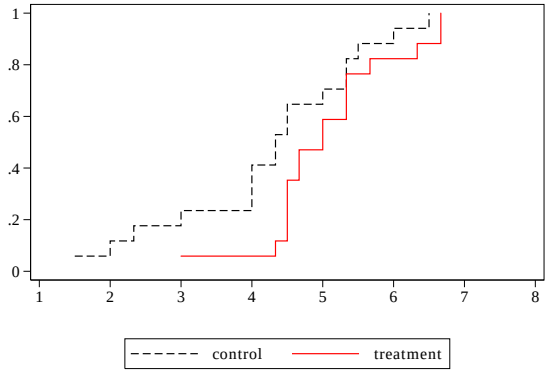
(b) Level of Emotional Adjustments, Females
$p = .017$



(c) Degree of Trust of Total Environment, Males
$p = .246$



(d) Degree of Trust of Total Environment, Females
$p = .026$



Notes: Each YRS item in these charts is an average over non-missing observations at ages 7, 8, and 9. Each item takes values from 1 to 7, with higher numbers corresponding to more socially-desirable behaviors like more learning or less stealing. "p" denotes p-values testing if the difference in means between the treatment and control groups is zero rather than a positive number.

E Identification and Parameter Restrictions

This section discusses the identifiability of the model used to generate the estimates reported in this paper. We specify our model in Section E.1 and discuss identifiability in Section E.2. In Section E.2, we present tests of overidentification which are used to check our model specification. We also show that the outcome decomposition for the effect of changes in skills on treatment effects, defined in equation (7) of the paper, is invariant to affine transformations of measures (Section (E.3)). For general discussions of identification in factor models, see Anderson and Rubin (1956) and Abbring and Heckman (2007).

E.1 Model Specification

In estimating outcome equation (5) of Section II.A, we recognize that skills are latent variables not directly observed but rather measured with error using multiple proxies. We use a *factor model* to estimate latent skills.

Factor analysis is a statistical method that explains the variability among observed measures in terms of latent skills (factors). It corrects for measurement error. It summarizes the information content of measures into a low-dimensional vector of skills (e.g. Wansbeek and Meijer (2000)). In this framework, skills $\boldsymbol{\theta}$ are called *factors* and measures are used to estimate factors through a set of linear equations called the *measurement system.* We assume in this paper that each measure is associated with at most one factor. A measurement system with this property is called a *dedicated measurement system.* More precisely, let the index set for measures associated with factor $j \in \mathcal{J}_p$ be $\mathcal{M}^j$. We denote the measures for factor j by $M^j_{m^j,d}$, where $m^j \in \mathcal{M}^j$, $d \in \{0,1\}$. Each factor j may be associated with a different number of measures. Henceforth we denote the vector of factors associated with the measured variables $(\theta^j_d : j \in \mathcal{J}_p), d \in \{0,1\}$ by $\boldsymbol{\theta}_d$.

Our model is as follows:

$$\text{The First Measure} : M^j_{1,d} = \nu^j_1 + \varphi^j_1 \theta^j_d + \eta^j_1,\ j \in \mathcal{J}_p \tag{E-1}$$

$$\text{Remaining Measures} : M^j_{m^j,d} = \nu^j_{m^j} + \varphi^j_{m^j} \theta^j_d + \eta^j_{m^j},\ j \in \mathcal{J}_p. \tag{E-2}$$

We distinguish the "first measure" from the "remaining measures," anticipating the normalizations required in factor analysis.

$$\text{Outcomes} : Y_d = \tau_d + \boldsymbol{\alpha}\boldsymbol{\theta}_d + \epsilon_d \tag{E-3}$$

$$\text{Factor Means} : E[\theta^j_d] = \mu^j_d,\ \forall\, j \in \mathcal{J}_p \tag{E-4}$$

$$\text{Factor Covariance} : \text{Var}[\boldsymbol{\theta}_d] = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_d}, \tag{E-5}$$

where $d \in \{0,1\}$, $m^j \in \mathcal{M}^j$, and $j \in \mathcal{J}_p$. We suppress the covariates $\boldsymbol{X}$ for notational simplicity. They are used in all empirical analyses. This convention is maintained throughout the analysis of this section. Equations (E-1) and (E-2) define our measurement system.

Parameters $\nu^j_{m^j}$ are measure-specific intercept terms. Parameters $\varphi^j_{m^j}$ are the factor loadings of the measurement system. Equation (E-3) defines the outcome equation. Parameter τ_d is an outcome-specific intercept term and parameters $\boldsymbol{\alpha} = (\alpha^j : j \in \mathcal{J}_p)$ are the outcome factor loadings. ϵ_d and $\eta^j_{m^j}$ are zero-mean error terms independent of $\boldsymbol{\theta}_d, d \in 0, 1$. Equations (E-4) and (E-5) define factor means and factor covariances.

E.2 Model Identification

We first establish conditions under which the model is identified.

Normalization: Standard identification of factor models requires fixing the location and the scale of factors (e.g. Anderson and Rubin (1956).) We set the location by fixing the intercepts of the first measure of each skill to zero, i.e. $\nu^j_1 = 0,\ j \in \mathcal{J}_p$, and we set the scale by fixing the factor loadings of the first measure of each skill to one, i.e. $\varphi^j_1 = 1,\ j \in \mathcal{J}_p$. We

show that decomposition (7) used in the text is invariant to the choice of the first measure, so long as $\varphi_1^j \neq 0$. By invariant we mean that even though factor loadings $\boldsymbol{\alpha}$ and skill treatment effects $E[\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0]$ may change when different normalizations are used, the values $\alpha^j E(\theta_1^j - \theta_0^j)$; are invariant for all $j \in \mathcal{J}_p$. Decomposition (7) is invariant to any affine transformations of measures (see section E.3 below).

Model identification is established in four steps. First, we identify the factor means μ_d^j. Second, we identify the factor loadings $\varphi_{m^j}^j$ for the measurement equation, the variances $\text{Var}(\eta_{m^j}^j)$ of the measurement system, and the factor covariance structure ($\boldsymbol{\Sigma}_{\boldsymbol{\theta}_d}$). Third, we identify the measure system intercepts $\nu_{m^j}^j$. Finally, we identify the factor loadings $\boldsymbol{\alpha}$ and intercept τ_d of the outcome equations. We now discuss these steps in the order given.

1. Factor Means We identify μ_1^j and μ_0^j from the expectation of the designated first measure for treatment and controls groups as

$$E(M_{1,d}^j) = \mu_d^j,\ j \in \mathcal{J}_p, d \in \{0,1\}. \tag{E-6}$$

2. Measurement Loadings From the covariance structure of the measurement system we identify the factor loadings of the measurement system (equation (E-7)), factor variances (equation (E-8)), variances of the measurement error term (equation (E-9)), and factor covariances (equation (E-10)). Thus

$$\varphi_{m^j}^j = \frac{\text{Cov}(M_{m^j,d}^j, M_{(m^j)',d}^j)}{\text{Cov}(M_{1,d}^j, M_{(m^j)',d}^j)} \quad \text{if } \text{Cov}(M_{1,d}^j, M_{(m^j)',d}^j) \neq 0, \tag{E-7}$$

$$\text{Var}(\theta_d^j) = \frac{\text{Cov}(M_{1,d}^j, M_{m^j,d}^j)}{\varphi_{m^j}^j} \quad \text{if } \varphi_{m^j}^j \neq 0, \tag{E-8}$$

$$\text{Var}(\eta_{m^j}^j) = \text{Var}(M_{m^j,d}^j) - [\varphi_{m^j}^j]^2 \,\text{Var}(\theta_d^j), \tag{E-9}$$

$$\text{Cov}(M_{1,d}^j, M_{1,d}^{j'}) = \text{Cov}(\theta_d^j, \theta_d^{j'}) \ \text{ for all } j, j' \in \mathcal{J}_p; j' \neq j. \tag{E-10}$$

A sufficient condition for identification in a dedicated factor model is the existence of three (or more) dedicated measures for each skill $j \in \mathcal{J}_p$ provided that all have a nonzero factor loading. For a general discussion, see Anderson and Rubin (1956). Notice that in equation (E-7) $\varphi^j_{m^j}$ might depend on $d \in \{0,1\}$, that is $\varphi^j_{m^j,d}$. In this case we must normalize $\varphi^j_{1,1} = \varphi^j_{1,0}$ to set a common scale across treatment and control groups. We test the hypothesis $H_0 : \varphi^j_{m^j,1} = \varphi^j_{m^j,0}, m_j \neq 1$, and we do not reject (see Table L.4 below).

3. Measurement Intercepts From the measurement equation

$$\nu^j_{m^j} = E(M^j_{m^j,d}) - \varphi^j_{m^j}\mu^j_d. \tag{E-11}$$

We can identify $\nu^j_{m^j}$, $m^j \in \mathcal{M}^j \setminus \{1\}$, $j \in \mathcal{J}_p$, since the factor loadings $\varphi^j_{m^j}$, $m^j \in \mathcal{M}^j$ and factor means μ^j_d for $j \in \mathcal{J}_p, d \in \{0,1\}$ are identified.

For much of our analysis we assume that the intercept $\nu^j_{m^j}$ for each component of each measurement equation does not depend on d. This assumption facilitates interpretability. If $\nu^j_{m^j}$ does not depend on d, then the treatment effect on measures, $E(M^j_{m^j,1}) - E(M^j_{m^j,0})$ operates solely through treatment effects on factor means, i.e. $\mu^j_1 - \mu^j_0$.

However, this condition is not strictly required. Model identification only requires intercept equality across treatment states for the first measure of each factor. Thus identification still holds if we allow all of the measurement intercepts to vary with treatment status indicator d except for the intercept of the designated first measure of each factor. We perform a robustness check by testing the equality of intercepts $H_0 : \nu^j_{m^j,1} = \nu^j_{m^j,0}$ for all measures except the designated first one. We do not reject the hypothesis of equality of intercepts for any factor (Table L.4).

4. Outcome Equation Suppose that $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_0$. Factor loadings for the outcome equation can be identified using the covariance between outcomes and the designated first measure of

each skill. The covariance between an outcome Y_d and the first measure of skill j, $M^j_{1,d}$, is

$$\operatorname{Cov}(Y_d, M^j_{1,d}) = \left(\alpha^j \operatorname{Var}(\theta^j_d) + \sum_{j' \in \mathcal{J}_p \setminus \{j\}} \alpha^{j'} \operatorname{Cov}(\theta^j_d, \theta^{j'}_d) \right). \tag{E-12}$$

Equation (E-12) can be represented in a more concise form. For notational brevity, stack the covariance of outcome Y_d across the first measures of all skills $j \in \mathcal{J}_p$ to obtain $\operatorname{Cov}(Y_d, M_{1,d}) = [\operatorname{Cov}(Y_d, M^j_{1,d}), j \in \mathcal{J}_p]$. Using this notation, we can represent the set of equations (E-12) for all factors $j \in \mathcal{J}_p$ by $\operatorname{Cov}(Y_d, M_{1,d}) = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_d} \boldsymbol{\alpha}$. Notice that $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_d}$ is identified from the argument of step 2. Therefore, $\boldsymbol{\alpha}$ is identified whenever $\det(\boldsymbol{\Sigma}_{\boldsymbol{\theta}_d}) \neq 0$.

Notice that it is straightforward to relax the assumption that $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_0$. We can allow the factor loadings of the outcome equation to depend on $d \in \{0, 1\}$. They can be identified through $\operatorname{Cov}(Y_d, M_{1,d}) = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_d} \boldsymbol{\alpha}_d$. We test if $H_0 : \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_0$ $m^j \in \mathcal{M}^j$, $j \in \mathcal{J}_p$, and we do not reject these hypotheses (see Tables L.2 and L.3). We interpret this result as evidence that the restricted specification of the outcome equation is valid. From $E(Y_d)$, we can identify τ_d because all the other parameters in this equation are identified.

E.3 Invariance to Affine Transformations of Measures

We now establish conditions under which outcome decomposition (7), relating treatment effects to experimentally induced changes in skills, is invariant to affine transforms of any measure of skill for any factor. Decomposition (7) assumes $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_0$. We also consider forming decompositions for the more general case where $\boldsymbol{\alpha}_1 \neq \boldsymbol{\alpha}_0$. We establish the invariance of (7) but not that of other terms in the decompositions that arise in the more general case. Throughout we assume autonomy of the measurement system so that intercepts and factor loadings are the same for treatments and controls for all measurement equations.

Before presenting a formal analysis, it is useful to present an intuition for its conclusions. Let $\tilde{M}^j_{m^j,d}$ be an affine transformation of the measure $M^j_{m^j,d}$, for some $j \in \mathcal{J}_p$ and $m^j \in \mathcal{M}^j$.

Specifically, define $\tilde{M}^j_{m^j,d}$ by:

$$\tilde{M}^j_{m^j,d} = aM^j_{m^j,d} + b \text{ such that } a \in \mathbb{R} \setminus \{0\}, b \in \mathbb{R}, \text{ and } d \in \{0,1\}, \text{ for all } j \in \mathcal{J}_p. \quad \text{(E-13)}$$

Let $\tilde{\varphi}^j_{m^j}, \tilde{\eta}^j_{m^j}, \tilde{\nu}^j_{m^j}$ be the factor loading, error term and intercept associated with the transformed measure $\tilde{M}^j_{m^j,d}, d \in \{0,1\}$. The key condition for the invariance of decomposition (7) to linear transformations of the different measures is that $\sum_{j \in \mathcal{J}_p} \alpha^j E(\theta^j_1 - \theta^j_0)$ be invariant.

We apply the same normalization to the transformed system as we do to the original system. Suppose that the measure transformed is a "first measure" so $m_j = 1$. Then, in the original system, $\nu^j_1 = 0$ and $\varphi^j_1 = 1$. Transformation (E-13) can be expressed as

$$\tilde{M}^j_{1,d} = b + a\theta^j_d + a\eta^j_1.$$

Applying the normalization rule to this equation defines factor $\tilde{\theta}_j = b + a\theta_j$, i.e. the scale and the location of the factor are changed, so that in the transformed system the intercept is 0 and the factor loading 1:

$$\tilde{M}^j_{1,d} = \tilde{\theta}^j_d + \tilde{\eta}^j_1$$

where $\tilde{\eta}^j_1 = a\eta^j_1$ is a rescaled mean zero error term. This transformation propogates through the entire system, where θ^j_d is replaced by $\tilde{\theta}^j_d$.

Notice that in decomposition (7), the induced shift in the mean of the factor is irrelevant. It differences out in the decomposition. The scale of θ^j is affected. The covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta_d}}$ is transformed to $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{d}}}$ where

$$\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{d}}} = \boldsymbol{I}_a \boldsymbol{\Sigma}_{\boldsymbol{\theta_d}} \boldsymbol{I}_a$$

where $\boldsymbol{I}_a$ is a square diagonal matrix of the same dimension as the number of measured factors and the j^{th} diagonal is a and the other elements are unity. From the analysis surrounding equation (E-12), the factor loading for the outcome function for the set of transformed first

measures, $\tilde{\boldsymbol{M}}_{\mathbf{1},\boldsymbol{d}} = \boldsymbol{M}_{1,d}\boldsymbol{I}_a$ is the solution to the system of equations

$$\mathrm{Cov}(Y_d, \tilde{\boldsymbol{M}}_{1,d}) = \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{d}}}\tilde{\boldsymbol{\alpha}}_d.$$

Thus

$$\begin{aligned}
\tilde{\boldsymbol{\alpha}}_{\boldsymbol{d}} &= \boldsymbol{\Sigma}^{-1}_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{d}}}\,\mathrm{Cov}(\boldsymbol{Y}_d, \tilde{\boldsymbol{M}}_{1,d}) \\
&= \boldsymbol{I}_a^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\boldsymbol{d}}}\boldsymbol{I}_a^{-1}\,\mathrm{Cov}(\boldsymbol{Y}_d, \tilde{\boldsymbol{M}}_{\mathbf{1},\boldsymbol{d}}) \\
&= \boldsymbol{I}_a^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\boldsymbol{d}}}\,\mathrm{Cov}(\boldsymbol{Y}_d, \boldsymbol{M}_{1,d}) \\
&= \boldsymbol{I}_a^{-1}\alpha_d.
\end{aligned}$$

Since $\tilde{\boldsymbol{\theta}}_{\boldsymbol{d}} = \boldsymbol{I}_a\boldsymbol{\theta}_{\boldsymbol{d}}$, it follows trivially that decomposition (7), $\boldsymbol{\alpha}'_{\boldsymbol{D}}(\boldsymbol{\theta}_{\mathbf{1}} - \boldsymbol{\theta}_{\mathbf{0}})$, is invariant to transformations.

Suppose next that the transformation is applied to any measure other than a first measure. Invoking the same kind of reasoning, it is evident that $\tilde{\boldsymbol{\theta}}_{\boldsymbol{d}} = \boldsymbol{\theta}_{\boldsymbol{d}}$ and $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{d}} = \boldsymbol{\alpha}_{\boldsymbol{d}}$. Thus the decomposition is invariant. Clearly, however, the intercept of the transformed measure becomes

$$\tilde{\nu}^j_{m_j} = b + a\nu^j_{m_j}$$

and the factor loading becomes

$$\tilde{\varphi}^j_{m_j} = \varphi^j_{m_j}a.$$

The preceding decomposition assumes that the outcome system is autonomous: $\boldsymbol{\alpha}_{\mathbf{0}} = \boldsymbol{\alpha}_{\mathbf{1}}$, and $\boldsymbol{\beta}_{\mathbf{0}} = \boldsymbol{\beta}_{\mathbf{1}}$. Suppose that $\boldsymbol{\alpha}_{\mathbf{1}} \neq \boldsymbol{\alpha}_{\mathbf{0}}$ but, to simplify the argument, we continue to assume that $\boldsymbol{\beta}_{\mathbf{0}} = \boldsymbol{\beta}_{\mathbf{1}}$. In this case

$$E(Y_1 - Y_0) = E(\boldsymbol{\alpha}'_{\mathbf{1}}\boldsymbol{\theta}_{\mathbf{1}} - \boldsymbol{\alpha}'_{\mathbf{0}}\boldsymbol{\theta}_{\mathbf{0}}).$$

In the general case, the decomposition is not unique due to a standard index number

problem. Using the notation $\Delta\boldsymbol{\alpha} = \boldsymbol{\alpha_1} - \boldsymbol{\alpha_0}$,

$$
\begin{aligned}
E(Y_1 - Y_0) = & \underbrace{\boldsymbol{\alpha_0'} E(\boldsymbol{\theta_1} - \boldsymbol{\theta_0})}_{\substack{\text{invariant to affine} \\ \text{transformations of} \\ \text{measures}}} + \underbrace{(\Delta\boldsymbol{\alpha})' E(\boldsymbol{\theta_1})}_{\substack{\text{non invariant to affine} \\ \text{transformations of} \\ \text{measures}}} \\
= & \underbrace{\boldsymbol{\alpha_1'} E(\boldsymbol{\theta_1} - \boldsymbol{\theta_0})}_{\substack{\text{invariant to affine} \\ \text{transformations of} \\ \text{measures}}} - \underbrace{(\Delta\boldsymbol{\alpha}) E(\boldsymbol{\theta_0})}_{\substack{\text{non-invariant to affine} \\ \text{transformations of} \\ \text{measures}}} .
\end{aligned}
$$

For any $\boldsymbol{\alpha^*}$ that is an affine transformation of (α_0, α_1)

$$
E(Y_1 - Y_0) = \underbrace{(\boldsymbol{\alpha^*}) E(\boldsymbol{\theta_1} - \boldsymbol{\theta_0})}_{\text{invariant to affine transformation}} + \underbrace{(\boldsymbol{\alpha_1} - \boldsymbol{\alpha^*}) E(\boldsymbol{\theta_1}) - (\boldsymbol{\alpha_0} - \boldsymbol{\alpha^*}) E(\boldsymbol{\theta_0})}_{\text{non-invariant to affine transformations}} .
$$

For all three decompositions, the term associated with the mean change in skills due to treatment is invariant to affine transformations. The proof follows the preceding reasoning. Any scaling of the factors is offset by the revised scaling of the factor loadings.

Notice, however that when $\boldsymbol{\alpha_1} \neq \boldsymbol{\alpha_0}$, we acquire terms in the level of the factors in constructing decompositions of treatment effects . For transformations to the first measure, the change in the *location* is shifted. Even though the scales of $(\Delta\boldsymbol{\alpha})$ and $E(\boldsymbol{\theta_d})$ offset, there is no compensating shift in the location of the factor. Thus the terms associated with the levels of the factor are not, in general invariant to affine transformations of first measures although the decompositions are invariant to monotonic transformations of any non-normalization measures. Obviously the point of evaluation against $E(\boldsymbol{\theta_1} - \boldsymbol{\theta_0})$ is evaluated depends on the choice of $\boldsymbol{\alpha_0}$, $\boldsymbol{\alpha_1}$, and $\boldsymbol{\alpha^*}$ if they differ. Heckman and Pinto (2012) generalize this result to general non-autonomous systems. The term associated with the change in $\boldsymbol{\theta}$ is invariant. The term associated with the changes in the function is not.

We now formally establish these results. It is enough to consider the transformation of one measure within group j for treatment category d. First, suppose that the transformation (E-13) is not applied to the first measure, that is, $m^j \neq 1$. In this case, $E(\theta_1^j - \theta_0^j)$; $j \in \mathcal{J}_p$ are invariant as they are identified through the first measure of each factor (Equation (E-6))

which is not changed. We can also show that the α^j, $j \in \mathcal{J}_p$, are invariant. We identify $\boldsymbol{\alpha} = [\alpha^j;\ j \in \mathcal{J}_p]$ through $\text{Cov}(Y_d, \boldsymbol{M}_{1,d}) = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_d}\boldsymbol{\alpha}$ (Equation (E-12)). Thereby it suffices to show that covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_d}$ is invariant under the linear transformation (E-13). But the covariance between the factors is identified through the first measure of each factor (equation (E-10)). And, according to equations (E-7)–(E-8), the variance of the factor j under transformation (E-13) is identified by:

$$\begin{aligned}
\frac{\text{Cov}(M^j_{1,d}, \tilde{M}^j_{m,d})\,\text{Cov}(M^j_{1,d}, M^j_{m',d})}{\text{Cov}(\tilde{M}^j_{m,d}, M^j_{m',d})} &= \frac{\text{Cov}(M^j_{1,d}, aM^j_{m,d})\,\text{Cov}(M^j_{1,d}, M^j_{m',d})}{\text{Cov}(aM^j_{m,d}, M^j_{m',d})} \qquad \text{by (E-13)} \\
&= \frac{a\,\text{Cov}(M^j_{1,d}, M^j_{m,d})\,\text{Cov}(M^j_{1,d}, M^j_{m',d})}{a\,\text{Cov}(M^j_{m,d}, M^j_{m',d})} \\
&= \frac{\text{Cov}(M^j_{1,d}, M^j_{m,d})\,\text{Cov}(M^j_{1,d}, M^j_{m',d})}{\text{Cov}(M^j_{m,d}, M^j_{m',d})} \\
&= \text{Var}(\theta^j_d),
\end{aligned}$$

so that the variance is unchanged. Hence $\boldsymbol{\alpha_d}$ is unchanged.

Now suppose that transformation (E-13) is applied to the first measure, $m^j = 1$. In this case, according to Equations (E-7)–(E-8), the new variance of factor j is given by:

$$\begin{aligned}
\frac{\text{Cov}(\tilde{M}^j_{1,d}, M^j_{m,d})\,\text{Cov}(\tilde{M}^j_{1,d}, M^j_{m',d})}{\text{Cov}(M^j_{m,d}, M^j_{m',d})} &= \frac{a\,\text{Cov}(M^j_{1,d}, M^j_{m,d})a\,\text{Cov}(M^j_{1,d}, M^j_{m',d})}{\text{Cov}(M^j_{m,d}, M^j_{m',d})} \\
&= a^2\,\text{Var}(\theta^j_d). \qquad \text{(E-14)}
\end{aligned}$$

According to Equation (E-10), the new covariance between factors j and j' is given by:

$$\begin{aligned}
\text{Cov}(\tilde{M}^j_{1,d}, M^{j'}_{1,d}) &= a\,\text{Cov}(M^j_{1,d}, M^{j'}_{1,d}) \\
&= a\,\text{Cov}(\theta^j_d, \theta^{j'}_d) \qquad \text{(E-15)}
\end{aligned}$$

Let $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}_d}$ be the new factor covariance matrix obtained under transformation (E-13). According to Equations (E-14)–(E-15), $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}_d} = \boldsymbol{I}_a\boldsymbol{\Sigma}_{\boldsymbol{\theta}_d}\boldsymbol{I}_a$, where, as before, $\boldsymbol{I}_a$ is a square diagonal

matrix whose j-th diagonal element is a and has ones for the remaining diagonal elements. By the same type of reasoning, we have that the covariance matrix $\mathrm{Cov}(Y_d, \boldsymbol{M}_{1,d})$ computed under the transformation is given by: $\mathrm{Cov}(Y_d, \tilde{\boldsymbol{M}}_{1,d}) = \boldsymbol{I}_a \mathrm{Cov}(Y_d, \boldsymbol{M}_{1,d})$. Let $\tilde{\boldsymbol{\alpha}}$ be the outcome factor loadings under transformation (E-13). Thus, by equation Equation (E-12),

$$\boldsymbol{I}_a \mathrm{Cov}(Y_d, \boldsymbol{M}_{1,d}) = \mathrm{Cov}(Y_d, \tilde{\boldsymbol{M}}_{1,d}) = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}_d} \tilde{\boldsymbol{\alpha}} = \boldsymbol{I}_a \boldsymbol{\Sigma}_{\boldsymbol{\theta}_d} \boldsymbol{I}_a \tilde{\boldsymbol{\alpha}} \tag{E-16}$$

and therefore $\tilde{\boldsymbol{\alpha}} = \boldsymbol{I}_a^{-1} \boldsymbol{\alpha}$. In other words, transformation (E-13) only modifies the j-th factor loading which is given by $\tilde{\alpha}^j = \frac{\alpha^j}{a}$.

Let the difference in factor means between treatment groups be $\Delta^{j'} = E(\theta_1^{j'} - \theta_0^{j'}), j' \in \mathcal{J}_p$, and let $\tilde{\Delta}^{j'}$ be the difference under transformation (E-13). According to Equation (E-6), transformation (E-13) only modifies the j-th difference in means which is given by $\tilde{\Delta}^j = a\Delta^j$ and thereby $\tilde{\alpha}^j \tilde{\Delta}^j = \alpha^j \Delta^j$. Thus $\tilde{\alpha}^{j'} \tilde{\Delta}^{j'} = \alpha^{j'} \Delta^{j'} = \alpha^{j'} E(\theta_1^{j'} - \theta_0^{j'})$ for all $j' \in \mathcal{J}_p$, as claimed. It is straightforward to establish that if $\boldsymbol{\alpha_1} \neq \boldsymbol{\alpha_0}$, the decomposition is, in general, not invariant to affine transformations, although the term associated with $E(\boldsymbol{\theta_1} - \boldsymbol{\theta_0})$ is. This holds for more general transformations than just the class of affine transformations, see (Heckman and Pinto, 2012).

F Correcting for Measurement Error Arising from Using Estimated Factor Scores

In this appendix we discuss the statistical properties of our three-step estimation procedure, which uses factor scores as regressors. The use of factor scores as regressors has a well-known drawback: due to measurement errors in the estimated factors, using factor scores as regressors produces biased estimates of the coefficients in the outcome equations, according to a standard errors-in-the-variables argument.

Bolck, Croon and Hagenaars (2004) show that a naive use of predicted latent scores as regressors generates biased estimators that understate the strength of the association between the outcomes and latent variables. A few methods are known to avoid these biases. Bolck, Croon and Hagenaars (2004) and Croon (2002) show that a simple correction of estimated parameters can eliminate this systematic bias. Following this line of research, Lu and Thomas (2008) present a correction framework, known as a "bias correction" approach. It is closely related to the traditional approach to solving errors-in-the-variables problems as described, for example, in Wansbeek and Meijer (2000). Skrondal and Laake (2001) and Lu and Thomas (2008) adopt an approach called "bias avoidance," which produces consistent estimators for OLS parameters in outcome equations by using a specific combination of regression factor scores for the explanatory latent variables and Bartlett (1937) scores for the response latent variables. We adapt the "bias correction" approach to accommodate two non-standard aspects of our model: (1) we estimate different measurement systems for the control group and for the treatment group; (2) each measurement system generates factor score predictors which are pooled to estimate a common outcome equation.

This appendix has two subsections. In Section F.1, we first discuss the statistical theory that supports the use of factor scores. In Section F.2, we explain how to correct the OLS regression to account for measurement errors in the factor scores.

F.1 Factor Scores

Our approach is based on a three-step procedure. We use a measurement system to evaluate factor scores $\boldsymbol{\theta}_S$, which, in turn, are used as covariates in outcome equations. Below is a description of the three steps.

1. First, a three-factor model is estimated. The vector of these factors for person i is denoted by $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_i^j : j \in \mathcal{J}_p)$.

2. Second, factor scores $\boldsymbol{\theta}_i$ are estimated for each participant i, based on the estimated parameters of the first step. We denote the resulting vector of factor scores by $\boldsymbol{\theta}_{S,i}$.

3. Finally, outcomes are regressed on the factor scores, identifying effects of factors on the outcome equations.

Let the measurement system for agent i, $i \in \{1, \ldots, N\}$ be written as:

$$\underbrace{\boldsymbol{M}_i}_{|\mathcal{M}|\times 1} = \underbrace{\boldsymbol{\varphi}}_{|\mathcal{M}|\times|p|} \underbrace{\boldsymbol{\theta}_i}_{|p|\times 1} + \underbrace{\boldsymbol{\eta}_i,}_{|\mathcal{M}|\times 1}$$

where $\boldsymbol{\varphi}$ represents a matrix of the factor loadings estimated in the first step and $\boldsymbol{M}_i$ is the vector of stacked measures for participant i with intercepts $\nu_{m^j}^j$ of Equation (8) removed. The dimension of each term is shown beneath it, with $\mathcal{M} = \cup_{j\in\mathcal{J}_p}\mathcal{M}^j$ being the union of all measure index sets. Let $\text{Cov}(\boldsymbol{\eta}_i, \boldsymbol{\eta}_i) = \boldsymbol{\Omega}$. We assume that the $(\boldsymbol{\theta}_i, \boldsymbol{\eta}_i)$ are independent across the participants. For simplicity, we assume that they are iid.[14] Let $\text{Cov}(\boldsymbol{M}_i, \boldsymbol{M}_i) = \boldsymbol{\Sigma}$, $\text{Cov}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i) = \boldsymbol{\Phi}$ and $\text{Cov}(\boldsymbol{\eta}_i, \boldsymbol{\eta}_i) = \boldsymbol{\Omega}$. Our estimation procedure produces consistent estimators of these covariance matrices and the parameters $\boldsymbol{\varphi}$.

We seek to estimate a vector of factor scores $\boldsymbol{\theta}_{S,i}$ that proxy the vector of latent skills $\boldsymbol{\theta}_i$ for each participant i. The most commonly used estimators of factor scores are based on a linear function of measures, that is, $\boldsymbol{\theta}_{S,i} = \boldsymbol{L}'\boldsymbol{M}_i$. Thurstone (1935) developed a linear estimator that minimizes the mean squared error (MSE) of the factor scores as predictors

[14] This is not strictly required but simplifies notation.

of the factors, but his estimator is biased. Linear unbiased estimators are obtained if the matrix relationship $\boldsymbol{L}'\boldsymbol{\varphi} = \boldsymbol{I}_{|\mathcal{J}|}$ is satisfied. Examples of this type of estimator are found in Bartlett (1937) and Horst (1965). His estimator is based on the restricted minimization of mean square error subject to $\boldsymbol{L}'\varphi = \boldsymbol{I}_{|\mathcal{J}|}$, which guarantees unbiasedness. His estimator is given by

$$\boldsymbol{L}^{B'} = (\boldsymbol{\varphi}'\boldsymbol{\Omega}^{-1}\boldsymbol{\varphi})^{-1}\boldsymbol{\varphi}'\boldsymbol{\Omega}^{-1}. \tag{F-1}$$

The factor score predictor is written as

$$\boldsymbol{\theta}_{S,i}^{B} = \boldsymbol{L}^{B'}\boldsymbol{M}_i = (\boldsymbol{\varphi}'\boldsymbol{\Omega}^{-1}\boldsymbol{\varphi})^{-1}\boldsymbol{\varphi}'\boldsymbol{\Omega}^{-1}\boldsymbol{M}_i. \tag{F-2}$$

Bartlett's estimator is a GLS procedure where measures are taken as dependent variables and factor loadings are treated as regressors. By the Gauss-Markov theorem, if the $\boldsymbol{\varphi}$ are known, the Bartlett GLS estimator is optimal and hence leads to the best linear unbiased predictor. Horst (1965) proposes a simpler OLS procedure that does not account for the heteroscedasticity of the error covariance matrix $\boldsymbol{\Omega}$. We adopt the Bartlett approach because of its more desirable statistical properties.[15]

F.2 Correcting for Estimation Error in the Factor Scores

Consider the model

$$Y_i = \boldsymbol{\alpha}\boldsymbol{\theta}_i + \boldsymbol{\gamma}\boldsymbol{Z}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, N. \tag{F-3}$$

The Covariance matrix of $(\boldsymbol{\theta}_i, \boldsymbol{Z}_i)$ is

$$\begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\theta}} & \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{Z}} \\ \boldsymbol{\Sigma}_{\boldsymbol{Z},\boldsymbol{\theta}} & \boldsymbol{\Sigma}_{\boldsymbol{Z},\boldsymbol{Z}} \end{pmatrix}.$$

[15] When $\boldsymbol{\varphi}$ is consistently estimated, we obtain a large sample version of the Guass-Markov theorem, and replace "unbiased" with "asymptomatically unbiased".

It is assumed that $\boldsymbol{\theta}_i$ is measured with error. Let $\boldsymbol{\theta}_{S,i}$ be a measure of $\boldsymbol{\theta}_i$, thus:

$$\boldsymbol{\theta}_{S,i} = \boldsymbol{\theta}_i + \boldsymbol{V}_i, \quad i = 1, \ldots, N;$$
$$(\boldsymbol{Z}_i, \boldsymbol{\theta}_i) \perp\!\!\!\perp \boldsymbol{V}_i, \quad E(\boldsymbol{V}_i) = 0, \quad \mathrm{Cov}(\boldsymbol{V}, \boldsymbol{V}) = \boldsymbol{\Sigma}_{\boldsymbol{VV}}.$$

We adopt the notation that $\boldsymbol{\Sigma}_{\boldsymbol{B},\boldsymbol{C}}$ is $\mathrm{Cov}(\boldsymbol{B}, \boldsymbol{C})$. Thus $\mathrm{Cov}(\boldsymbol{\theta}_{S,i}, \boldsymbol{\theta}_{S,i})$ is $\boldsymbol{\Sigma}_{\boldsymbol{\theta_S},\boldsymbol{\theta_S}}$.

We assume that the $(\boldsymbol{\theta}_i, \boldsymbol{Z}_i, \boldsymbol{\epsilon}_i)$ are iid, but our analysis applies under much weaker conditions. For example, it is enough to require the vector to be independent, but not necessarily identically distributed across observations. Notice that we do not assume that $\boldsymbol{\theta}_i \perp\!\!\!\perp \boldsymbol{Z}_i$ as in traditional factor analysis. However, we do assume that $(\boldsymbol{\theta}_i, \boldsymbol{Z}_i) \perp\!\!\!\perp \boldsymbol{\epsilon}_i$ and $E(\boldsymbol{\epsilon}_i) = 0$ where "$\perp\!\!\!\perp$" denotes independence.

By a standard argument, using Y_i in place of $\boldsymbol{\theta}_{S,i}$, we obtain

$$Y_i = \boldsymbol{\alpha}\boldsymbol{\theta}_{S,i} + \boldsymbol{\gamma}\boldsymbol{Z}_i + \boldsymbol{\epsilon}_i - \boldsymbol{\alpha}\boldsymbol{V}_i. \tag{F-4}$$

The OLS estimator is inconsistent:

$$\mathrm{plim} \begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} = \begin{pmatrix} \mathrm{Cov}(\boldsymbol{\theta}_S, \boldsymbol{\theta}_S) & \mathrm{Cov}(\boldsymbol{\theta}_S, \boldsymbol{Z}) \\ \mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{\theta}_S) & \mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{Z}) \end{pmatrix}^{-1} \begin{pmatrix} \mathrm{Cov}(\boldsymbol{\theta}, \boldsymbol{\theta}) & \mathrm{Cov}(\boldsymbol{\theta}, \boldsymbol{Z}) \\ \mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{\theta}) & \mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{Z}) \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\gamma} \end{pmatrix}.$$

Observe that $\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{Z}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_S,\boldsymbol{Z}}$ as a consequence of our assumptions. In this notation,

$$\mathrm{plim} \begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} = \underbrace{\begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{V},\boldsymbol{V}} & \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{Z}} \\ \boldsymbol{\Sigma}_{\boldsymbol{Z},\boldsymbol{\theta}} & \boldsymbol{\Sigma}_{\boldsymbol{Z},\boldsymbol{Z}} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\theta}} & \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{Z}} \\ \boldsymbol{\Sigma}_{\boldsymbol{Z},\boldsymbol{\theta}} & \boldsymbol{\Sigma}_{\boldsymbol{Z},\boldsymbol{Z}} \end{pmatrix}}_{\boldsymbol{A}} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\gamma} \end{pmatrix}. \tag{F-5}$$

This is the usual attenuation formula.

Notice that from estimates of the measurement system, we can identify $\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\theta}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\theta},Z}$, $\boldsymbol{\Sigma}_{\boldsymbol{V},\boldsymbol{V}}$, and hence all components of $\boldsymbol{A}$. Thus, if we pre-multiply the least squares estimator by $\boldsymbol{A}^{-1}$,

we obtain:

$$\text{plim}\, \boldsymbol{A}^{-1} \begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\gamma} \end{pmatrix}.$$

This is called "Croon's method" in psychometrics (Croon, 2002).

In our application, there are two groups corresponding to $D = 0$ and $D = 1$ (control and treatment, respectively). We allow $\boldsymbol{\theta}_i$ to vary by treatment status. Our method assumes that treatment only operates through shifting the distribution of $\boldsymbol{\theta}$. We do not normalize the means of $\boldsymbol{\theta}$ (or $\boldsymbol{Z}$) to be zero.

G Sufficient Conditions Guaranteeing Unbiased Estimates of Factor Loadings of Outcome Equations

In this section we examine conditions under which the estimators of the outcome factor loadings are unbiased. The key assumption in this appendix is that the latent skills are independent of the $\boldsymbol{X}$, and that the latent skills are measured without error. As before, we use $\mathcal{J}$ for the index of set of skills. We use $\mathcal{J}_p \subset \mathcal{J}$ for the subset of measured skills.

Equation (5) describes an outcome of interest for a treatment d as a linear function of an intercept τ_d, skills $(\theta_d^j; j \in \mathcal{J}_p)$ and pre-program variables $\boldsymbol{X}$:

$$Y_d = \tau_d + \sum_{j \in \mathcal{J}_p} \alpha_d^j \theta_d^j + \boldsymbol{\beta X} + \epsilon_d,\ d \in \{0,1\}. \tag{G-1}$$

The intercept term τ_d is $\tau_d = \kappa_d + \sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha_d^j E(\theta_d^j)$. The error term ϵ_d is given by $\epsilon_d = \tilde{\epsilon}_d + \sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha_d^j (\theta_d^j - E(\theta_d^j))$. We also assume that errors $\tilde{\epsilon}_d$ are mean zero i.i.d. random variable such that $\tilde{\epsilon}_d \perp\!\!\!\perp ((\theta_d^j; j \in \mathcal{J}), \boldsymbol{X})$ for $d \in \{0.1\}$. From the independence of $\boldsymbol{X}$ and the latent skills,

$$E(\epsilon_d | \boldsymbol{X}) = E(\epsilon_d) = 0 \text{ for } d \in \{0,1\}.$$

Equation (G-1) can be used to represent the outcome equation as a standard linear regression equation comprising both treatment groups (Equation (6)):

$$\begin{aligned} Y &= D \underbrace{(\tau_1 + \sum_{j \in \mathcal{J}_p} \alpha^j \theta_1^j + \boldsymbol{\beta X} + \epsilon_1)}_{Y_1} + (1 - D) \underbrace{(\tau_0 + \sum_{j \in \mathcal{J}_p} \alpha^j \theta_0^j + \boldsymbol{\beta X} + \epsilon_0)}_{Y_0} \\ &= \tau_0 + \tau D + \sum_{j \in \mathcal{J}_p} \alpha^j \theta^j + \boldsymbol{\beta X} + \epsilon, \end{aligned}$$

where $\tau = \tau_1 - \tau_0$ is the contribution of unmeasured variables to mean treatment effects, $\epsilon = D\epsilon_1 + (1 - D)\epsilon_0$ is a mean-zero error term, and $\theta^j = D\theta_1^j + (1 - D)\theta_0^j$, $\forall\, j \in \mathcal{J}_p$ denotes the skills that we can measure.

Our goal is examine whether the least squares estimators α^j, $\forall j \in \mathcal{J}_p$ are unbiased when the measured skills $(\theta_d^j; j \in \mathcal{J}_p)$ are independent of unmeasured ones $(\theta_d^j; j \in \mathcal{J} \setminus \mathcal{J}_p)$.

Lemma G-1. *If skills and treatment status indicators are independent conditional on pre-program variables* $\boldsymbol{X}$*, i.e.*

$$((\theta_1^j; j \in \mathcal{J}), (\theta_0^j; j \in \mathcal{J})) \perp\!\!\!\perp D|\boldsymbol{X}, \text{ (Randomization Assumption)} \tag{G-2}$$

and measured and unmeasured skills are independent, conditional on $\boldsymbol{X}$*,*

$$((\theta_d^j; j \in \mathcal{J}_p) \perp\!\!\!\perp (\theta_d^j; j \in \mathcal{J} \setminus \mathcal{J}_p))|\boldsymbol{X} \text{ for } d \in \{0,1\}, \text{ (Skill Independence)} \tag{G-3}$$

then linear regression (6) *generates unbiased estimate of* $(\alpha^j; j \in \mathcal{J}_p)$.

Proof. It suffices to prove that $E(\epsilon|\boldsymbol{X}, D, (\theta^j; j \in \mathcal{J}_p)) = 0$. But $\epsilon = D\epsilon_1 + (1-D)\epsilon_0$. By independence assumption (G-2) and the definition of $\tilde{\epsilon}_d$; $d\{0,1\}$, it follows that:

$$E(\epsilon|\boldsymbol{X}, D) = E(\epsilon|\boldsymbol{X}, D = d) = E(\epsilon_d|\boldsymbol{X}) = 0 \text{ for } d \in \{0,1\}. \tag{G-4}$$

Thus, it is enough to show that assumptions (G-2) and (G-3) imply that

$$(\theta^j; j \in \mathcal{J}_p) \perp\!\!\!\perp \epsilon|\boldsymbol{X}, D.$$

Conditioning on $D = d$ reduces the preceding expression to

$$(\theta_d^j; j \in \mathcal{J}_p) \perp\!\!\!\perp \epsilon_d|\boldsymbol{X}.$$

Recall that ϵ_d is a function of $(\theta_d^j; j \in \mathcal{J} \setminus \mathcal{J}_p))$, which are independent of $(\theta_d^j; j \in \mathcal{J}_p)$ conditioned on $\boldsymbol{X}$ by assumption (G-3).

□

H Exploratory Factor Analysis

This appendix supplements the discussion on exploratory factor analysis in Section III. We review the background of factor rotation, define direct quartimin rotation, and establish how to extract a low-dimensional vector of latent factors that are proxied by multiple psychological measures. We perform a standard exploratory factor analysis (e.g., Gorsuch, 2003; Thompson, 2004) on the Perry PBI and YRS psychological measures that were described in Web Appendixes C and D.

H.1 Factor Rotation

We briefly review some basic aspects of a standard factor model in order to introduce the concept of factor rotation. We then explain the direct quartimin oblique rotation which is the rotation method used in our exploratory factor analysis.

A standard linear factor model is defined by:

$$\boldsymbol{M} = \boldsymbol{\nu} + \boldsymbol{\varphi}\boldsymbol{\theta} + \boldsymbol{\eta},$$

where $\boldsymbol{\theta}$ is a $|\mathcal{J}|$-dimensional vector of latent factors, $\boldsymbol{M}$ is an $|\mathcal{M}|$-dimensional vector of measures, $\boldsymbol{\nu}$ is an $|\mathcal{M}|$-dimensional vector of intercepts, and $\boldsymbol{\eta}$ is an $|\mathcal{M}|$-dimensional vector of error terms assumed to be independent of each other and of factors $\boldsymbol{\theta}$. $\boldsymbol{\varphi}$ is the vector of factor loadings and has dimension $|\mathcal{M}| \times |\mathcal{J}|$. The identification of the mean of the factors is trivial and does not affect the rotation analysis (see Anderson and Rubin (1956) for details). Thus we assume that the means of the factors, measures and error terms are zero.

A major goal of factor analysis is to explain the variability of a set of variables called *measures* into two components: a *common portion* explained by a set of latent variables called *factors*, and a *unique portion* that is due to an idiosyncratic variation particular to each measure. More specifically, factor analysis decomposes the covariance matrix of measures into the sum of a covariance matrix explained by factors and an error term covariance that is

not explained by factors. We denote the covariance of a random vector X by $\sigma_X \equiv Var(X)$. Thus the assumptions made so far can be summarized in the following equations:

$$\boldsymbol{\Sigma_M} = \underbrace{\boldsymbol{\varphi \Sigma_\theta \varphi'}}_{\text{common portion}} + \underbrace{\boldsymbol{\Sigma_\eta}}_{\text{unique portion}}, \tag{H-1}$$

where $\boldsymbol{\Sigma_\eta}$ is diagonal.

Indeterminacy There is an inherent indeterminacy in factor models, as Equation (H-1) can be rewritten as

$$\boldsymbol{\Sigma_M} = (\boldsymbol{\varphi U^{-1}})(\boldsymbol{U \Sigma_\theta U'})(\boldsymbol{\varphi U^{-1}})' + \boldsymbol{\Sigma_\eta}, \tag{H-2}$$

for any non-singular $|\mathcal{J}|$-dimensional square matrix $\boldsymbol{U}$. We refer to $\boldsymbol{U}$ as a rotation matrix and it can be interpreted as a linear rotation of factor axes that does not change the share of explained variation of measures. Rotation is an important tool for facilitating interpretation of the data. As explained below, factor axis can be rotated to achieve a simpler factor structure, which, in turn, adds to the interpretability of the factors.

Interpretation A simple linear regression model is useful for understanding how a rotation can improve factor interpretability. Suppose an analyst wants to evaluate the impact of verbal and motor cognitive abilities on childhood development. The analyst can perform a linear regression that uses the sum and difference of these abilities as covariates. In this case, the interpretation of the parameters associated with these two covariates is unclear. Instead, the analyst can recover the actual measures of verbal and motor cognitive abilities through linear operations. Using the actual skill measures instead of their sum and difference improves the interpretability of the estimated parameters. In both models, the analyst explains the same fraction of the variation of the target outcomes.

Simplicity As mentioned, a rotation can generate a simplified factor structure which increases the factor interpretability. A notion of factor simplicity was first proposed by Thur-

stone (1947). He suggests five properties that a simple matrix of factor loadings $\boldsymbol{\varphi}$ should have:

1. Each row contains at least one zero (i.e. there is no measure that is loaded on all factors);
2. Each column should have the same number of zeros as there are factors;
3. For any pair of factors, there are some variables with zero loadings on one factor and large loadings on the other factor;
4. For any pair of factors, there is a sizable proportion of zero loadings;
5. For any pair of factors, there is only a small number of large loadings.

While Thurstone's 1947 suggestions are useful to clarify the concept of simplicity, they lack mathematical precision. In particular, we cannot compare the simplicity of matrices of factor loadings that differ in more than one property. The literature on factor analysis has coped with this lack of formalism by providing a range of *simplicity criteria*, which are weighting functions that summarize the simplicity characteristics of a factor loading matrix into a single value. (See Jennrich, 2006)

Types of Rotations A rotation is usually computed by the matrix $\boldsymbol{U}$ that maximizes a simplicity criterion Q associated with a matrix of factor loadings $\boldsymbol{\varphi}$. Namely,

$$\boldsymbol{U} = \text{argmax}_{\tilde{\mathbf{U}} \in \mathcal{U}} \, Q(\boldsymbol{\varphi} \tilde{\boldsymbol{U}}^{-1}) \tag{H-3}$$

where $Q(\cdot)$ is a simplicity criteria and $\mathcal{U}$ is the set over which maximization is performed. We can impose rotation properties on the matrix U, such as invertibility, by addressing restrictions on the set $\mathcal{U}$. Another use of the set $\mathcal{U}$ is to generate *orthogonal* and *oblique* rotations. While orthogonal rotation imposes that new axes are orthogonal to each other, the oblique rotation relaxes this constraint. In other words, oblique rotation allows factors to

be correlated. In order to retain the factors normalized to their original length, the rotation matrix $\boldsymbol{U}$ has to be constrained so that $\text{diag}(\boldsymbol{U}'\boldsymbol{U}) = \boldsymbol{I}$, where $\text{diag}(\boldsymbol{A})$ denotes the diagonal matrix of $\boldsymbol{A}$ and $\boldsymbol{I}$ is the identity matrix (e.g., see Mulaik (1972) p. 308).

Quartimin rotation Simplicity is key in factor rotation. Most of the rotation criteria are based on the Crawford-Ferguson family (Crawford and Ferguson, 1970) of simplicity measures. The rotation criteria is defined as a weighted sum of the row (variable) and column (factor) simplicity inline with the Thurstone (1947) notions of simplicity (Browne, 2001). Specifically, let the matrix $\boldsymbol{A}$ be defined by $\boldsymbol{A} \equiv A_{j,k}\ ;\ j \in \{1,\ldots,|\mathcal{M}|\}, k \in \{1,\ldots,|\mathcal{J}|\}$, then the measure of simplicity is defined by:

$$Q(\boldsymbol{A}) = -\left((1-\gamma)\underbrace{\left(\sum_{j=1}^{|\mathcal{M}|}\sum_{k=1}^{|\mathcal{J}|}\sum_{l\neq k,l=1}^{|\mathcal{J}|} A_{j,k}^2 A_{j,l}^2\right)}_{\text{Row Complexity}} + \gamma\underbrace{\left(\sum_{k=1}^{|\mathcal{J}|}\sum_{j=1}^{|\mathcal{M}|}\sum_{l\neq j,l=1}^{|\mathcal{M}|} A_{j,k}^2 A_{l,k}^2\right)}_{\text{Column Complexity}}\right) \tag{H-4}$$

The *quartimin rotation* criterion was first developed by Jennrich and Sampson (1966), and it belongs to a family of oblique rotations that use the simplicity criteria proposed by Carroll (1953). Its formula is defined by setting γ in Equation (H-4) to zero. The quartimin rotation focuses on reducing the weight on row/variable complexity in order to obtain a perfect cluster configuration (Carroll, 1953).

It is intuitive why quartimin leads to simple structure of loadings like the one reported in Table H.2. In order to obtain small row complexity (see the first term in formula (H-4)), we need to have only one loading per row large, while all others close to zero. If all others were exactly zero, then row complexity would be zero. If more than one loading per row is large, the criterion (H-4) penalizes us by producing large row complexity, since the multiplication of two large squared numbers is a large number.

More succinctly, the quartimin simplicity criteria can be written as:

$$Q_q(\boldsymbol{A}) = -\text{trace}((\boldsymbol{A}\cdot\boldsymbol{A})(\boldsymbol{A}\cdot\boldsymbol{A})\boldsymbol{N})$$

where $\boldsymbol{A}$ is a target matrix, $(\boldsymbol{A} \cdot \boldsymbol{A})$ denotes a element-wise product and $\boldsymbol{N}$ is a square matrix with zeros on the diagonal and ones elsewhere. The quartimin rotation for the matrix of factor loadings $\boldsymbol{\varphi}$ is given by the matrix $\mathbf{U}$ that maximizes the following equation:

$$\begin{aligned} \boldsymbol{U} &= \operatorname{argmax} Q_q(\boldsymbol{\varphi}\tilde{\boldsymbol{U}}^{-1}) \\ &\text{s.t. } \tilde{\boldsymbol{U}} \text{ is invertible and } \operatorname{diag}(\boldsymbol{U}'\boldsymbol{U}) = \boldsymbol{I} \end{aligned}$$

H.2 Exploratory Factor Analysis

Exploratory Factor Analysis seeks *dedicated measures* of each factor, i.e., measures that proxy a single factor.[16] We search for dedicated measures using Exploratory Factor Analysis (EFA) with direct quartimin rotation. The method identifies blocks of measures that are highly loaded on one factor and negligibly loaded on other factors after a direct quartimin rotation. We exclude items that are weakly associated with factors.[17] We also exclude those items that are not clearly associated with one and only one particular factor, since they cannot serve as dedicated measures of any of the factors that we can account for in the model.[18]

Before searching for dedicated measures based on the EFA with quartimin method, we establish the number of factors to extract. A variety of criteria are offered in the literature (Gorsuch, 2003; Thompson, 2004; Zwick and Velicer, 1986). We use three separate procedures (the scree test (Cattell, 1966), Onatski's test (Onatski, 2009), and Horn's test (Horn, 1965)).[19] The scree test, Horn's test, and Onatski's test point to three factors for females and to a range from two to four factors for males. Both the scree test and Horn's test applied

[16] Factors based on dedicated measures are easily interpretable and not restricted to be orthogonal (see Section III of the paper).

[17] More specifically, we do not retain measures that do not have loadings at least .6 or higher for at least one gender (*the weak loading problem*).

[18] Namely, we do not retain measures that have at least two loadings greater than .4 (*the cross-loading problem*).

[19] Another rule, the Guttman-Kaiser rule, overestimates the number of factors (Zwick and Velicer, 1986) and so results based on this procedure are not very informative (≤ 9 factors).

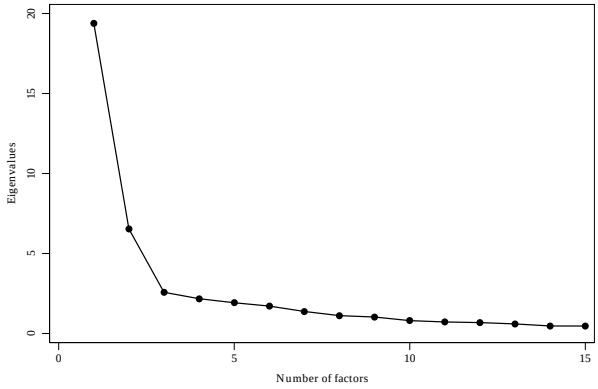
to the pooled sample of males and females suggest three factors, while Onatski's procedure fails to converge (see Figure H.1 and Table H.1). Overall, these results point to three factors as the most likely solution.

Table H.2 shows factor loadings for the final exploratory factor model after direct quartimin rotation.[20] Loadings in bold are substantially larger than other loadings for the same item. Moreover, the bolded loadings are always statistically significant, while the unbolded ones are generally not. Thus in our application EFA produces sensible results.

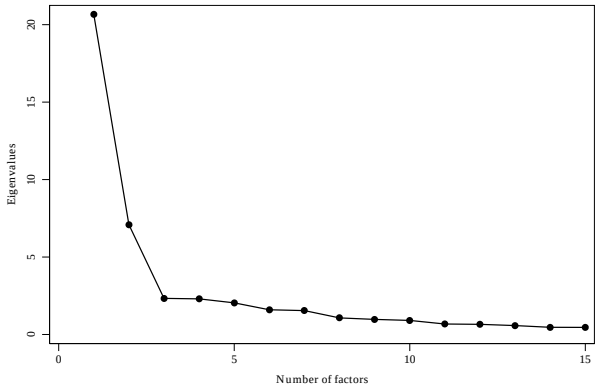
[20]We find that other widely recognized oblique rotations, such as geomin, lead to similar results and the same choice of measures as quartimin (see Table L.7 of the Web Appendix). This is in line with the literature showing that widely recognized methods produce similar results (Fabrigar et al., 1999).

Figure H.1: Scree Plots for All 46 Items

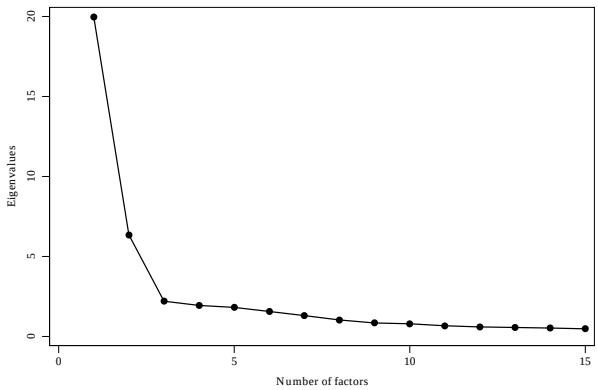
(a) Males



(b) Females



(c) Overall



Notes: See Tables C.1 and D.1 of Web Appendices C and D for a list of the 43 PBI and YRS items. See Web Appendix B for the three Stanford-Binet tests given at ages 7, 8, and 9.

Table H.1: Results of Procedures Estimating the Number of Factors Using All 46 Items[(a)]

Procedure	Males	Females	Pooled
Scree[(b)]	3	3	3
Horn[(c)]	4	3	3
Onatski[(d), (e)]	2	3	–[(f)]

[(a)] See Tables C.1 and D.1 of Web Appendices C and D for a list of the 43 PBI and YRS items. See Web Appendix B for the three Stanford-Binet tests given at ages 7, 8, and 9.
[(b)] Scree test by Cattell (1966). See Figure H.1 for scree plots.
[(c)] Horn's (1965) parallel analysis procedure.
[(d)] We apply Onatski's (2009) procedure at the 10% level of significance for a minimum of two factors and a maximum of five factors (we choose a minimum of two since we expect at least to have cognitive and personality factors). Onatski (2009) warns that the asymptotic approximation may be poor in a case like ours, where sample size is small and the number of measures is low.
[(e)] The Guttman-Kaiser-rule (Guttman, 1954; Kaiser, 1960, 1961) excludes factors that clearly have little explanatory power, but often overestimates the number of informative factors (Zwick and Velicer, 1986). In our application, it produced an upper bound of 7–9.
[(f)] Onatski's algorithm does not converge to any number in the range from two to five.

Table H.2: Factor Loadings of a Three-Factor Model After Oblique Rotation

	Males						Females						Pooled					
	Cognition	Standard Error	Externalizing Behavior	Standard Error	Academic Motivation	Standard Error	Cognition	Standard Error	Externalizing Behavior	Standard Error	Academic Motivation	Standard Error	Cognition	Standard Error	Externalizing Behavior	Standard Error	Academic Motivation	Standard Error
Cognition																		
Stanford Binet, age 7	**.666**	*(.099)*	-.030	*(.099)*	.123	*(.116)*	**.877**	*(.092)*	-.104	*(.063)*	.120	*(.108)*	**.783**	*(.059)*	-.052	*(.051)*	.097	*(.071)*
Stanford Binet, age 8	**.700**	*(.086)*	-.104	*(.084)*	.222	*(.106)*	**.846**	*(.081)*	.012	*(.075)*	.052	*(.093)*	**.805**	*(.055)*	-.049	*(.047)*	.112	*(.067)*
Stanford Binet, age 9	**.925**	*(.063)*	.070	*(.047)*	.008	*(.049)*	**.885**	*(.077)*	.070	*(.072)*	-.054	*(.076)*	**.899**	*(.045)*	.067	*(.043)*	-.035	*(.044)*
Externalizing Behavior																		
Disrupts classroom procedures	-.084	*(.072)*	**.782**	*(.065)*	.176	*(.088)*	-.094	*(.092)*	**.824**	*(.070)*	.193	*(.106)*	-.077	*(.056)*	**.802**	*(.047)*	.176	*(.069)*
Swears or uses obscene words	-.154	*(.076)*	**.732**	*(.075)*	.235	*(.094)*	.023	*(.122)*	**.698**	*(.093)*	.235	*(.130)*	-.075	*(.066)*	**.724**	*(.059)*	.218	*(.080)*
Steals	-.010	*(.134)*	**.371**	*(.134)*	.119	*(.150)*	-.007	*(.116)*	**.736**	*(.087)*	.156	*(.123)*	.037	*(.103)*	**.486**	*(.093)*	.107	*(.113)*
Lying or cheating	-.155	*(.095)*	**.569**	*(.101)*	.332	*(.115)*	-.045	*(.102)*	**.783**	*(.075)*	.179	*(.112)*	-.086	*(.074)*	**.639**	*(.070)*	.279	*(.089)*
Influences others toward troublemaking	-.037	*(.058)*	**.927**	*(.043)*	-.028	*(.066)*	-.021	*(.081)*	**.905**	*(.047)*	.019	*(.084)*	-.027	*(.046)*	**.932**	*(.031)*	-.027	*(.052)*
Aggressive toward peers	.260	*(.077)*	**.841**	*(.065)*	-.145	*(.071)*	.107	*(.084)*	**.892**	*(.067)*	-.303	*(.091)*	.182	*(.064)*	**.839**	*(.049)*	-.176	*(.061)*
Teases or provokes students	.053	*(.078)*	**.834**	*(.063)*	-.059	*(.086)*	.085	*(.148)*	**.691**	*(.106)*	-.188	*(.154)*	.040	*(.075)*	**.782**	*(.057)*	-.112	*(.079)*
Academic Motivation																		
Shows Initiative	.076	*(.051)*	-.065	*(.042)*	**.910**	*(.047)*	.042	*(.086)*	.002	*(.058)*	**.899**	*(.068)*	.037	*(.042)*	-.064	*(.030)*	**.917**	*(.037)*
Alert and interested in school work	.082	*(.051)*	.069	*(.055)*	**.895**	*(.054)*	.202	*(.112)*	.162	*(.070)*	**.779**	*(.106)*	.100	*(.046)*	.098	*(.045)*	**.866**	*(.049)*
Hesitant to try, or gives up easily	.049	*(.088)*	.195	*(.100)*	**.664**	*(.093)*	.273	*(.150)*	.090	*(.115)*	**.555**	*(.139)*	.121	*(.078)*	.175	*(.077)*	**.631**	*(.080)*
Sample size			59						37						96			

Notes: Factor loadings based on the exploratory factor analysis with direct quartimin rotation (Jennrich and Sampson, 1966) are shown. Maximum likelihood asymptotic standard errors are in parentheses. Factor loadings relating factors to corresponding potential dedicated measures are in bold.

I Notes on Power

The small sample size of the Perry Study may call into question the power of hypothesis tests performed on it. We show that this concern is overstated. Following the standard literature on power analysis, we compute the minimum effect size that is likely to be detected in a sample of the size of the Perry study. We compute power and significance level. Following standard conventions, we assume that treatment and control outcomes are normally distributed with different means but with equal variances.

Statistical power is the probability that a test rejects the null hypothesis when it is false. Effect size is the standardized mean difference between treatments and controls. Lower levels of the effect size required to reject a false null implies greater statistical power.

Power depends on the choice of the critical value, set by defining the significance level. The statistical power of a test depends on the variance in the sample, the sample size, and the specific alternative hypothesis against which the null hypothesis is being contrasted.

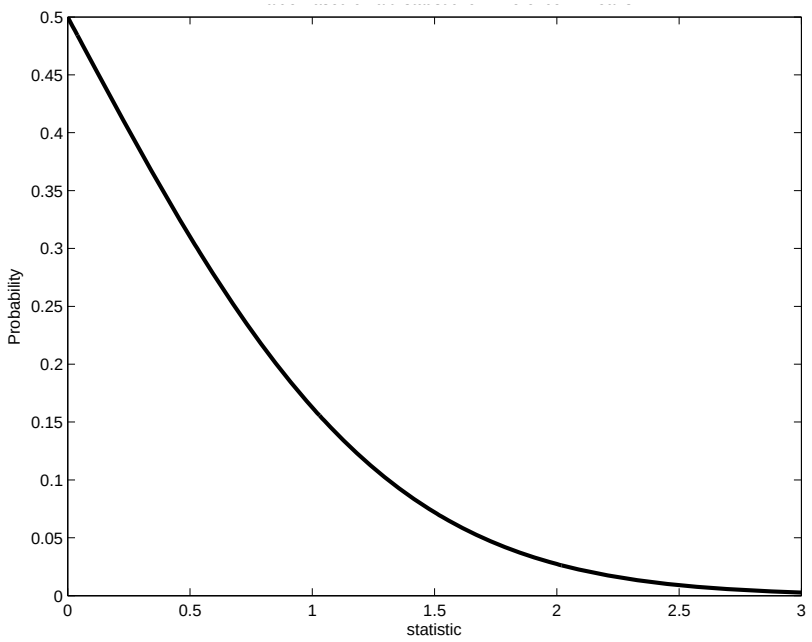
There are 123 participants in the Perry sample. Our analyses are gender-specific. Thus we adopt a sample size of 60 in our calculations. We assume that both treatment and control groups have 30 observations. Table I.1 gives the critical values (effect sizes) for significance levels of 10% and 5% and 1% for the t-statistic. Figure I.1 shows the p-values for testing the one-sided single hypothesis of no treatment effects under different values of the t-statistics associated with the difference in means between treatment groups. The t-statistic is a sufficient statistic to compute the p-values. Figure I.2 shows the p-values for testing the one-sided single hypothesis of no treatment effects under different values of sample variance and for different values of the difference in means across treatment groups.

Table I.1: Critical Values

Significance Levels	t-stat	Effect size
10%	1.31	0.34
5%	1.70	0.44
1%	2.46	0.63

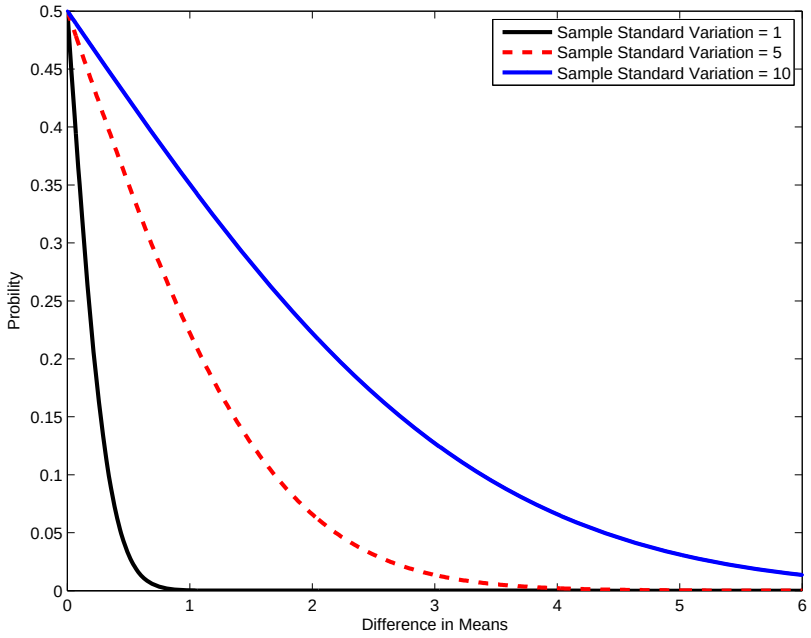
This table shows the critical values for testing the one-sided single hypothesis of no treatment effects. Treatment and control outcomes are normal with different means but with equal variance. The Perry sample consists of 123 participants, but the analyses are gender-specific. Thus we adopt a sample size of 60 in our calculations. We assume that both treatment and control groups have 30 observations each. The first column in the table gives significance levels. The second column gives the critical values of one-sided t-statistics for the significance levels of 10% and 5% and 1%. The last column gives critical values for the effect size for the significance levels of 10% and 5% and 1% given our sample.

Figure I.1: p-values for the t-statistic of the Difference in Means



This figure shows graphically the critical values for testing the one-sided single hypothesis of no treatment effects. We adopt the conventional approach in which the treatment and control outcomes are normally distributed with different means but equal variances. The Perry sample consists of 123 participants, but the analyses are gender-specific. Thus we adopt a sample size of 60 in our calculations. We assume that both treatment and control groups have 30 observations each. The figure shows the p-values associated with t-statistic critical values.

Figure I.2: p-values for the Difference in Means and Sampling Variation



This figure shows graphically the critical values for testing the one-sided single hypothesis of no treatment effects. We adopt the conventional approach in which the treatment and control outcomes are normally distributed with different means but equal variances. The Perry sample consists of 123 participants, but the analyses are gender-specific. Thus we adopt a sample size of 60 in our calculations. We assume that both treatment and control groups have 30 observations each. The figure shows how sample variation affect p-values for a range of the difference in means across treatment groups.

Table I.2 shows the effect sizes and t-statistics for a range of values of statistical power and significance levels. Specifically, if the data generating process is such that the effect size is 0.4 (first line and first column of panel (a)), then a testing procedure that adopts a significance level of 10% would correctly reject the null hypothesis of no-treatment effect at a 60% rate. According to panel (a) Table I.2, for achieving a power level of 80%, we need an effect size of .55 for a significance level of 10%. We need an effect size of 0.65 for a significance level of 5% and an effect size of 0.85 for a significance level of 1%.

Table I.3 shows the statistical power for testing the one-sided single hypothesis of no treatment effects. We compute power based on the effect sizes of the Perry outcomes as presented in Table 1. The Perry sample consists of 123 participants, but our analyses are gender-specific. Thus we adopt a sample size of 72 for males and 51 for females in our calculations. We adopt the conventional approach in which the treatment and control outcomes are normally distributed with different means but with equal variance. We adopt the traditional significance levels of 10% and 5% and 1%. A total of 75% of male outcomes have statistical power beyond 50% at a 10% significance level. For females, this percentage is 85%. Half of the outcomes have statistical power beyond 60% at a 10% significance level for both genders.

Table I.2: Power Critical Values

(a) Power for the Effect Size of the Difference in Means

	Power for Effect Size		
Significace Level	60%	70%	80%
10%	0.40	0.48	0.55
5%	0.50	0.58	0.65
1%	0.70	0.77	0.85

(b) Power for the t-statistic of the Difference in Means

	Power for t-statistic		
Significace Level	60%	70%	80%
10%	1.56	1.84	2.16
5%	1.95	2.23	2.55
1%	2.71	2.99	3.31

This tables show the statistical power associated with testing the one-sided hypothesis of no treatment effects. Treatment and control outcomes are normally distributed with different means but have equal variances. The Perry sample consists 123 participants. The analyses are gender-specific. Thus we adopt a sample size of 60 in our calculations. We assume that both treatment and control groups have 30 observations each. Panel (a) presents the statistical power associated with the the effect sizes reported in the table. Panel (b) gives statistical power associated with different t-statistics of the difference in means. The first column of the tables gives the significance levels of 10% and 5% and 1%. The second column the statistic needed for a power of 60% for each significant level. Specifically, if the data generating process is such that the effect size is 0.4 (first line and first column of panel (a)), then an inference that adopts a significance level of 10% would correctly reject the null hypothesis of no-treatment effect at a 60% rate. The remaining columns provide the level of statistics needed for powers of 70% and 80% respectively.

Table I.3: Power for Perry Outcome (Males and Females)

Variable	Age	Treatment Effect			Statistical Power		
					Significance Level		
		Effect	Effect Size	p-value	0.10	0.05	0.01
A. Males							
CAT total at age 14, end of grade 8	14	0.566 *	0.652	***(0.060)***	0.93	0.86	0.65
# of misdemeanor arrests, age 27	27	-1.21 **	-0.363	***(0.036)***	0.60	0.45	0.21
# of felony arrests, age 27	27	-1.12	-0.324	(0.101)	0.53	0.39	0.16
# of adult arrests (misd.+fel.), age 27	27	-2.33 **	-0.402	***(0.024)***	0.66	0.52	0.26
Monthly income, age 27	27	0.876 **	0.607	***(0.018)***	0.90	0.82	0.58
Use tobacco, age 27	27	-0.119 *	-0.236	***(0.093)***	0.39	0.26	0.09
# of misdemeanor arrests, age 40	40	-3.13 **	-0.372	***(0.039)***	0.61	0.47	0.22
# of felony arrests, age 40	40	-1.14 *	-0.266	***(0.092)***	0.44	0.30	0.11
# of adult arrests (misd.+fel.), age 40	40	-4.26 **	-0.373	***(0.041)***	0.61	0.47	0.22
# of lifetime arrests, age 40	40	-4.20 *	-0.346	***(0.053)***	0.57	0.42	0.19
Employed, age 40	40	0.200 **	0.394	***(0.024)***	0.65	0.50	0.25
Sample		72					
B. Females							
CAT total, age 8	8	0.565 *	0.614	***(0.062)***	0.82	0.70	0.43
CAT total, age 14	14	0.806 **	0.909	***(0.014)***	0.98	0.94	0.81
Any special education, age 14	14	-0.262 ***	-0.514	***(0.009)***	0.71	0.57	0.30
Mentally impaired at least once, age 19	19	-0.280 **	-0.569	***(0.029)***	0.77	0.65	0.37
# of misdemeanor violent crimes, age 27	27	-0.423 **	-0.292	***(0.032)***	0.41	0.27	0.10
# of felony arrests, age 27	27	-0.269 **	-0.325	***(0.021)***	0.45	0.31	0.12
Jobless for more than 1 year, age 27	27	-0.292 **	-0.573	***(0.038)***	0.78	0.65	0.38
Ever tried drugs other than alcohol or weed, age 27	27	-0.227 **	-0.530	***(0.045)***	0.73	0.59	0.32
# of misdemeanor violent crimes, age 40	40	-0.537 **	-0.364	***(0.016)***	0.51	0.36	0.15
# of felony arrests, age 40	40	-0.383 **	-0.425	***(0.028)***	0.59	0.45	0.20
# of lifetime violent crimes, age 40	40	-0.574 **	-0.384	***(0.019)***	0.54	0.39	0.16
Months in all marriages, age 40	40	39.6 *	0.539	***(0.076)***	0.74	0.61	0.33
Sample		51					

This table shows the statistical power for testing the one-sided single hypothesis of no treatment effects. The first column of the table describes the male and female outcomes. The next three columns of the table are taken from Table 1. The reported effect is the difference in means between treatment and control groups. The stars denote statistical significance: *** - 1 percent level, ** - 5 percent level, * - 10 percent level. The effect size is the ratio of the effect to the standard deviation of the control group. The fourth column provides the one-sided single hypothesis p-value associated with the test of no treatment effects. The remaining three columns of the tables report the statistical power for testing the one-sided single hypothesis of no treatment effects for significance levels of 10%, 5% and 1% respectively. We adopt the conventional approach in which the treatment and control outcomes are normally distributed with different means but with equal variance. The Perry sample consists of 123 participants, but our analyses are gender-specific. Thus we adopt a sample size of 72 for males and 51 for females in our calculations.

Multiple measures on the same, or similar outcomes, and covariates $\boldsymbol{X}$ enhance the power of the Perry study by (a) controlling (or eliminating) the effect of measurement error, and (b) reducing residual variance. Our use of factors controls for measurement error and presents low dimensional summaries of the data that conserve on degrees of freedom.

J Assumptions Required for Testing H_0: $plim\,\hat{\boldsymbol{\alpha}}_1 = plim\,\hat{\boldsymbol{\alpha}}_{\mathbf{0}}$.

In this section we examine minimal conditions for identifying the coefficients of the measured skills in the outcome equation in the presence of unmeasured skills. We use $\mathcal{J}$ for an index set of skills. We use $\mathcal{J}_p \subset \mathcal{J}$ for the subset of measured skills. As in the text, our model for the outcome equation is:

$$Y_d = \kappa_d + \sum_{j \in \mathcal{J}} \alpha_d^j \theta_d^j + \boldsymbol{\beta}_d \boldsymbol{X} + \tilde{\epsilon}_d, \;\; d \in \{0,1\},$$

where κ_d is an intercept, $(\alpha_d^j; j \in \mathcal{J})$ are factor loadings and $\boldsymbol{\beta}_d$ are $|\boldsymbol{X}|$-dimensional vectors of parameters. Error term $\tilde{\epsilon}_d$ is a zero-mean i.i.d. random variable assumed to be independent of regressors $(\theta_d^j; j \in \mathcal{J})$ and $\boldsymbol{X}$. We abstract from measurement error in the measured latent skills.

The Perry analysts collected a rich array of measures of cognitive and personality skills. However, it is likely that there are skills that they did not measure. Notationally, let $\mathcal{J}_p \subseteq \mathcal{J}$ be the index set of measured skills. Rewrite equation (2) for potential outcome Y_d as:

$$\begin{aligned}
Y_d &= \kappa_d + \sum_{j \in \mathcal{J}} \alpha_d^j \theta_d^j + \boldsymbol{\beta}_d \boldsymbol{X} + \tilde{\epsilon}_d \\
&= \kappa_d + \underbrace{\sum_{j \in \mathcal{J}_p} \alpha_d^j \theta_d^j}_{\substack{\text{effect from skills} \\ \text{that we measure}}} + \underbrace{\sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha^j \theta_d^j}_{\substack{\text{effect from skills that} \\ \text{we do not measure}}} + \boldsymbol{\beta}_d \boldsymbol{X} + \tilde{\epsilon}_d \\
&= \underbrace{\kappa_d + \sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha_d^j E(\theta_d^j)}_{\text{new intercept}} + \underbrace{\sum_{j \in \mathcal{J}_p} \alpha_d^j \theta_d^j}_{\substack{\text{effect from skills} \\ \text{that we measure}}} + \underbrace{\sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha_d^j (\theta_d^j - E(\theta_d^j))}_{\substack{\text{effect from skills that} \\ \text{we do not measure}}} + \boldsymbol{\beta}_d \boldsymbol{X} + \tilde{\epsilon}_d, \\
&= \underbrace{\tau_d}_{\text{new intercept}} + \underbrace{\sum_{j \in \mathcal{J}_p} \alpha_d^j \theta_d^j}_{\substack{\text{effect from skills} \\ \text{that we measure}}} + \boldsymbol{\beta}_d \boldsymbol{X} + \underbrace{\sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha_d^j (\theta_d^j - E(\theta_d^j)) + \tilde{\epsilon}_d}_{\text{new error term} = \epsilon_d}
\end{aligned} \tag{J-1}$$

where $d \in \{0,1\}$, $\tau_d = \kappa_d + \sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha_d^j E(\theta_d^j)$.Any differences in the error terms between

treatment and control groups can be attributed to differences in unmeasured skills. Thus we assume, without any loss of generality, that $\tilde{\epsilon}_1 \overset{dist}{=} \tilde{\epsilon}_0$, where $\overset{dist}{=}$ means equality in distribution.

It is easy to see that if unmeasured skills are independent of measures skills, namely,

$$(\theta_d^j;\ j \in \mathcal{J} \setminus \mathcal{J}_p) \perp\!\!\!\perp (\theta_d^j;\ j \in \mathcal{J}_p) | \boldsymbol{X};\ d \in \{0,1\},$$

then the regression:

$$Y_d = \tau_d + \sum_{j \in \mathcal{J}_p} \alpha_d^j \theta_d^j + \boldsymbol{\beta}_d \boldsymbol{X} + \epsilon_d, \tag{J-2}$$

produces unbiased estimates of parameter $(\alpha_d^j;\ j \in \mathcal{J}_p);\ d \in \{0,1\}$. Indeed error term ϵ_d in equation (J-2) is given by

$$\epsilon_d = \tilde{\epsilon}_d + \sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha_d^j (\theta_d^j - E(\theta_d^j))$$

which are independent of $(\theta_d^j;\ j \in \mathcal{J}_p)$ conditional on $\boldsymbol{X}$ under the assumption that skills are independent.

Now suppose instead of invoking the independence of skills assumption for both groups, we only assume it for the control group, and assume

$$(\theta_0^j;\ j \in \mathcal{J} \setminus \mathcal{J}_p) \perp\!\!\!\perp (\theta_0^j;\ j \in \mathcal{J}_p) | \boldsymbol{X}.$$

Moreover, suppose we also assume that $\alpha_1^j = \alpha_0^j;\ j \in \mathcal{J}$, so the outcome factor loadings for both treatment and control groups are the same. In this setup, the regression

$$Y_0 = \tau_0 + \sum_{j \in \mathcal{J}_p} \alpha^j \theta_0^j + \boldsymbol{\beta}_0 \boldsymbol{X} + \epsilon_0, \tag{J-3}$$

produces unbiased estimates of $(\alpha^j;\ j \in \mathcal{J}_p)$. Now consider the regression

$$Y_1 = \tau_1 + \sum_{j \in \mathcal{J}_p} \alpha^j \theta_1^j + \boldsymbol{\beta}_1 \boldsymbol{X} + \epsilon_1,$$

This regression produces unbiased estimates of $(\alpha^j;\ j \in \mathcal{J}_p)$ if:

$$(\theta_1^j;\ j \in \mathcal{J} \setminus \mathcal{J}_p) \perp\!\!\!\perp (\theta_1^j;\ j \in \mathcal{J}_p)|\boldsymbol{X}, \tag{J-4}$$

or alternatively,

$$(\theta_1^j - \theta_0^j;\ j \in \mathcal{J} \setminus \mathcal{J}_p) \perp\!\!\!\perp (\theta_1^j - \theta_0^j;\ j \in \mathcal{J}_p)|\boldsymbol{X}. \tag{J-5}$$

Thus, under this new set of assumptions, testing $H_0 : plim\ \hat{\boldsymbol{\alpha}}_1 = plim\ \hat{\boldsymbol{\alpha}}_0$, where $(\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_0)$ are estimates of $(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_0)$, is translated into testing the independence relations of equations (J-4)–(J-5).

K Decompositions Based on Simple Averages of Measures

This appendix presents decompositions of treatment effects using simple averages of measures.[21] We compare decompositions derived from simple averages with the decompositions derived from factor analysis. Estimates based on simple averages may suffer from attenuation bias, induced by uncorrected measurement error, although averaging goes part way in eliminating this bias. In addition, if the factor model is correct, and the factor loadings across measures are unequal, an unweighted average of the measures is a biased estimate of the factor.

The analysis in this appendix abstracts from an important practical problem. We have 43 psychological measures. We could use all 43 measures in our analysis or form averages of various subsets of these measures. These considerations pose a substantial model selection problem. We avoid this problem by taking the clusters of measures identified through EFA and forming simple averages of them—not accounting for measurement error or differential weighting of the measures that might be indicated by the factor analysis. We also present experiments that substitute CAT for IQ. Borghans et al. (2011) show that achievement tests capture different traits than IQ tests. (See also Heckman and Kautz, 2012.)

K.1 Empirical Results

We present estimates of models that are directly comparable to the model explicated in the text, with the only difference being that they are based on simple averages instead of factors. Results based on simple averages are similar to the results based on factors but are generally smaller in absolute value because of the attenuation bias, as demonstrated in Figure 8 of the main paper.

Tables K.1 and K.2 present estimates of models for males and females for the case when

[21] We use equally weighted averages.

$\alpha_1 = \alpha_0$, which is maintained in the main analysis in this paper. The tables show results of four one-index models as well as results of two three-index models. The one-index models explain outcomes based on just one of the four indices representing IQ, CAT, Externalizing Behavior, and Academic Motivation averaged over ages 7, 8, and 9. The first three-factor model is based on three factors: IQ, Externalizing Behavior, and Academic Motivation. The second three-factor model is similar, but uses CAT instead of IQ as a measure of cognition.

For males, the one-index models that regress the outcome on the single index show no effects of IQ, CAT, and Academic Motivation on outcomes. However, for a number of outcomes we can see strong and statistically significant or borderline statistically significant effects of Externalizing Behavior (see Table K.1). Three-index models show similar results: no effect of Cognition and Academic Motivation, but strong and statistically significant or borderline statistically significant effects of Externalizing Behavior for a number of outcomes. The estimated effects of the Externalizing Behavior in three-index model are generally several percentage points smaller than those for the one-factor model. For females, one-index models show strong estimates of effects for IQ, CAT, and Externalizing Behavior for a number of outcomes. For three-index models, CAT and Externalizing Behavior remain strong predictors.

Tables K.3 and K.4 use the same list of variables and in the same fashion as Tables K.1 and K.2 but for models with unrestricted coefficients ($\alpha_1 \neq \alpha_0$). Comparing Tables K.3 and K.4 to Tables K.1 and K.2 we can see that results of restricted and unrestricted models are close, which corroborates the tests reported in the text.

Table K.1: Restricted Decompositions ($\alpha_1 = \alpha_0$): Males

		Models Using One Index at a Time[a]				Models Using Three Indicies Simultaneously[b]							
						IQ and Noncog. Skills				CAT and Noncog. Skills			
Outcome	Statistic	IQ	CAT	Externalizing Behavior	Academic Motivation	IQ	Externalizing Behavior	Academic Motivation	Total Explained	CAT	Externalizing Behavior	Academic Motivation	Total Explained
CAT total at age 14, end of grade 8 (+)	effect	10%	20%	30%	6%	8%	2%	4%	14%	17%	5%	1%	23%
	p-value	.110	.179	.163	.176	.130	.462	.159	***.086***	.165	.383	.274	.153
# of misdemeanor arrests, age 27 (-)	effect	3%	7%	23%	2%	2%	20%	0%	22%	4%	19%	0%	23%
	p-value	.364	.305	***.080***	.431	.380	***.088***	.487	.119	.348	***.080***	.520	.117
# of felony arrests, age 27 (-)	effect	5%	10%	40%	3%	3%	35%	0%	38%	5%	35%	0%	40%
	p-value	.376	.292	.119	.411	.380	.118	.559	.150	.345	.124	.547	.141
# of adult arrests (misd.+fel.), age 27 (-)	effect	4%	8%	30%	3%	2%	26%	0%	28%	5%	25%	0%	30%
	p-value	.347	.287	***.084***	.420	.386	***.083***	.538	.105	.338	***.082***	.529	.103
Monthly income, age 27 (+)	effect	2%	5%	18%	2%	2%	13%	1%	16%	3%	13%	0%	16%
	p-value	.388	.427	.188	.587	.480	.226	.559	.341	.485	.218	.543	.332
Use tobacco, age 27 (-)	effect	4%	4%	54%	3%	3%	52%	-1%	55%	-6%	57%	1%	52%
	p-value	.425	.475	.187	.536	.454	.170	.534	.197	.557	.183	.536	.212
# of misdemeanor arrests, age 40 (-)	effect	2%	5%	16%	2%	1%	12%	1%	14%	3%	12%	0%	15%
	p-value	.396	.325	***.089***	.437	.423	.183	.468	.202	.385	.170	.474	.181
# of felony arrests, age 40 (-)	effect	6%	13%	54%	4%	4%	47%	0%	52%	8%	48%	-1%	55%
	p-value	.379	.325	.183	.431	.401	.180	.525	.188	.342	.194	.552	.205
# of adult arrests (misd.+fel.), age 40 (-)	effect	3%	7%	24%	2%	2%	20%	1%	22%	4%	19%	0%	23%
	p-value	.392	.306	***.091***	.443	.417	.109	.493	.139	.365	.105	.519	.118
# of lifetime arrests, age 40 (-)	effect	3%	8%	29%	3%	2%	25%	0%	28%	5%	25%	0%	29%
	p-value	.372	.320	.107	.428	.410	.111	.491	.137	.367	.112	.507	.128
Employed, age 40 (+)	effect	3%	2%	14%	1%	3%	13%	-1%	16%	-1%	15%	0%	14%
	p-value	.361	.405	.182	.499	.365	.172	.547	.206	.507	.183	.536	.225

Notes: Percentages of the treatment effect explained by indices of traits are shown. Regression coefficients for the treatment and control groups are restricted to be the same. One-sided p-values are based on 1000 bootstrap draws. p-values below 10% are in bold. Signs (+) and (-) represent the sign of the total treatment effect, which is shown in Table 1 of the text. Indices of Externalizing Behavior and Academic Motivation constructed in this paper are defined in Table 2 of the text. Indices of IQ and CAT are averages over IQ and CAT measures at ages 7, 8, and 9. [a] Outcomes are regressed on only one index. [b] Outcomes are regressed on a vector of indices. Two versions of this model are estimated. In one model, Cognition is IQ, while in another model it is CAT.

Table K.2: Restricted Decompositions ($\alpha_1 = \alpha_0$): Females

		Models Using One Index at a Time[a]				Models Using Three Indicies Simultaneously[b]							
						IQ and Noncog. Traits				CAT amd Noncog. Traits			
Outcome	Statistic	IQ	CAT	Externalizing Behavior	Academic Motivation	IQ	Externalizing Behavior	Academic Motivation	Total Explained	CAT	Externalizing Behavior	Academic Motivation	Total Explained
CAT total, age 8 (+)	effect	51%	-	29%	41%	24%	8%	34%	66%	-	-	-	-
	p- value	***.052***	-	***.055***	.145	.107	.354	.169	***.075***	-	-	-	-
CAT total, age 14 (+)	effect	30%	58%	37%	25%	10%	3%	21%	34%	95%	9%	2%	105%
	p- value	***.032***	***.010***	***.015***	.147	.167	.436	.129	.110	***.008***	.188	.406	***.008***
Any special education, age 14 (-)	effect	15%	33%	15%	9%	10%	11%	4%	26%	53%	-2%	3%	54%
	p- value	.140	***.022***	.179	.171	.235	.267	.311	.134	***.015***	.628	.328	***.028***
Mentally impaired at least once, age 19 (+)	effect	23%	32%	-4%	14%	16%	-13%	11%	14%	45%	13%	-8%	49%
	p- value	.103	***.035***	.609	.364	.186	.812	.338	.464	***.028***	.223	.761	***.024***
# of misdemeanor violent crimes, age 27 (-)	effect	-7%	4%	35%	2%	-8%	36%	0%	28%	28%	-11%	4%	20%
	p- value	.842	.369	***.072***	.317	.787	***.028***	.554	.114	.128	.813	.406	.318
# of felony arrests, age 27 (+)	effect	-10%	11%	36%	-1%	-7%	38%	-3%	28%	9%	37%	-5%	41%
	p- value	.755	.304	***.083***	.626	.610	***.061***	.741	.107	.355	***.031***	.734	.121
Jobless for more than 1 year, age 27 (-)	effect	8%	45%	21%	23%	-20%	0%	29%	9%	39%	40%	-18%	60%
	p- value	.334	***.064***	.185	.223	.790	.512	.233	.449	.198	***.071***	.790	***.096***
Ever tried drugs other than alcohol or weed, age 27 (-)	effect	-2%	-8%	24%	-10%	12%	30%	-15%	27%	10%	4%	19%	33%
	p- value	.618	.635	.122	.697	.153	***.078***	.702	.142	.363	.422	.262	.227
# of misdemeanor violent crimes, age 40 (-)	effect	-3%	11%	37%	4%	-5%	36%	1%	33%	30%	28%	-20%	37%
	p- value	.660	.261	***.076***	.287	.689	***.043***	.469	.102	***.082***	***.089***	.748	.141
# of felony arrests, age 40 (-)	effect	-2%	16%	35%	4%	-4%	35%	1%	32%	15%	37%	-5%	47%
	p- value	.567	.143	***.017***	.351	.601	***.017***	.533	***.044***	.273	***.054***	.675	.119
# of lifetime violent crimes, age 40 (-)	effect	-5%	10%	37%	2%	-5%	37%	0%	31%	27%	36%	-10%	54%
	p- value	.766	.265	***.023***	.359	.702	***.015***	.593	***.069***	.183	***.018***	.721	***.047***
Months in all marriages, age 40 (+)	effect	35%	37%	36%	15%	31%	32%	1%	64%	20%	38%	-9%	49%
	p- value	.110	***.091***	***.091***	.210	.137	.112	.466	***.081***	.109	***.022***	.827	***.057***

Notes: Percentages of the treatment effect explained by indices of traits are shown. Regression coefficients for the treatment and control groups are restricted to be the same. One-sided p-values are based on 1000 bootstrap draws. p-values below 10% are in bold. Signs (+) and (-) represent the sign of the total treatment effect, which is shown in Table 1 of the text. Indices of Externalizing Behavior and Academic Motivation constructed in this paper are defined in Table 2 of the text. Indices of IQ and CAT are averages over IQ and CAT measures at ages 7, 8, and 9. [a]Outcomes are regressed on only one index. [b]Outcomes are regressed on a vector of indices. Two versions of this model are estimated. In one model, Cognition is IQ, while in another model it is CAT. Signs "-" denote omitted estimates of models that make no sense since they have the early CAT on both right-hand and left-hand sides.

Table K.3: Unrestricted Decompositions ($\alpha_1 \neq \alpha_0$): Males

		Models Using One Index at a Time[a]				Models Using Three Indicies Simultaneously[b]							
						IQ and Noncog. Skills				CAT and Noncog. Skills			
Outcome	Statistic	IQ	CAT	Externalizing Behavior	Academic Motivation	IQ	Externalizing Behavior	Academic Motivation	Total Explained	CAT	Externalizing Behavior	Academic Motivation	Total Explained
CAT total at age 14, end of grade 8 (+)	effect	10%	21%	29%	6%	8%	5%	4%	17%	15%	6%	2%	23%
	p- value	.128	.172	.160	.176	.140	.342	.176	.108	.192	.323	.208	.160
# of misdemeanor arrests, age 27 (-)	effect	3%	6%	22%	2%	1%	18%	1%	19%	4%	16%	0%	20%
	p- value	.363	.294	***.064***	.457	.463	***.070***	.500	.132	.378	***.070***	.528	.136
# of felony arrests, age 27 (-)	effect	4%	9%	37%	3%	2%	28%	1%	31%	6%	30%	0%	35%
	p- value	.389	.293	***.056***	.428	.431	***.072***	.561	.133	.385	***.072***	.589	.131
# of adult arrests (misd.+fel.), age 27 (-)	effect	3%	7%	28%	3%	2%	22%	1%	24%	4%	22%	0%	26%
	p- value	.383	.289	***.061***	.426	.429	***.061***	.510	.121	.367	***.061***	.573	.115
Monthly income, age 27 (+)	effect	2%	5%	24%	2%	2%	18%	1%	21%	6%	18%	0%	23%
	p- value	.410	.422	.173	.582	.439	.189	.546	.273	.457	.171	.481	.244
Use tobacco, age 27 (-)	effect	4%	4%	57%	3%	3%	52%	-1%	55%	-4%	60%	1%	56%
	p- value	.416	.439	.119	.572	.460	.123	.504	.151	.537	.124	.533	.168
# of misdemeanor arrests, age 40 (-)	effect	2%	6%	15%	2%	1%	11%	1%	12%	3%	9%	0%	13%
	p- value	.381	.301	***.100***	.432	.491	.184	.468	.228	.396	.193	.492	.219
# of felony arrests, age 40 (-)	effect	6%	13%	53%	4%	4%	42%	0%	46%	9%	43%	-1%	51%
	p- value	.365	.294	***.056***	.421	.427	***.063***	.542	.112	.358	***.061***	.573	.104
# of adult arrests (misd.+fel.), age 40 (-)	effect	3%	7%	23%	2%	1%	17%	1%	19%	4%	16%	0%	21%
	p- value	.386	.296	***.063***	.451	.461	***.098***	.498	.166	.365	***.087***	.531	.147
# of lifetime arrests, age 40 (-)	effect	3%	8%	28%	3%	2%	22%	1%	25%	5%	21%	0%	26%
	p- value	.387	.299	***.070***	.444	.464	***.071***	.493	.128	.389	***.084***	.534	.130
Employed, age 40 (+)	effect	3%	0%	13%	1%	2%	9%	0%	10%	-1%	11%	0%	10%
	p- value	.357	.449	.200	.481	.398	.247	.547	.271	.510	.213	.554	.303

Notes: Percentages of the treatment effect explained by indices of traits are shown. No equality between coefficients of treatment and control groups is imposed. Decompositions are evaluated at the average level of the coefficients estimated for the treatment and control groups. One-sided p-values are based on 1000 bootstrap draws. p-values below 10% are in bold. Signs (+) and (-) represent the sign of the total treatment effect, which is shown in Table 1 of the text. Indices of Externalizing Behavior and Academic Motivation constructed in this paper are defined in Table 2 of the text. Indices of IQ and CAT are averages over IQ and CAT measures at ages 7, 8, and 9. $^{(a)}$A part of the treatment effect explained by changes in the index. $^{(b)}$A part of the treatment effect explained by changes in the coefficient.

Table K.4: Unrestricted Decompositions ($\alpha_1 \neq \alpha_0$): Females

		Models Using One Index at a Time[a]				Models Using Three Indicies Simultaneously[b]							
						IQ and Noncog. Traits				CAT amd Noncog. Traits			
Outcome	Statistic	IQ	CAT	Externalizing Behavior	Academic Motivation	IQ	Externalizing Behavior	Academic Motivation	Total Explained	CAT	Externalizing Behavior	Academic Motivation	Total Explained
CAT total, age 8 (+)	effect	52%	-	29%	42%	5%	11%	39%	55%	-	-	-	-
	p-value	***.032***	-	***.048***	.170	.438	.259	.181	.138	-	-	-	-
CAT total, age 14 (+)	effect	28%	54%	35%	25%	0%	2%	24%	26%	41%	-3%	7%	45%
	p-value	***.021***	***.005***	***.012***	.121	.530	.467	.133	.181	***.016***	.649	.225	***.038***
Any special education, age 14 (-)	effect	18%	34%	17%	9%	10%	12%	5%	27%	53%	13%	-10%	56%
	p-value	***.098***	***.009***	.126	.175	.259	.197	.297	.107	***.015***	.173	.816	***.014***
Mentally impaired at least once, age 19 (+)	effect	34%	36%	-6%	13%	21%	-18%	10%	12%	33%	-16%	2%	19%
	p-value	***.074***	***.025***	.663	.354	.245	.912	.356	.548	.134	.942	.435	.380
# of misdemeanor violent crimes, age 27 (-)	effect	-9%	6%	33%	1%	-20%	35%	2%	17%	8%	35%	-6%	38%
	p-value	.818	.390	***.008***	.376	.886	***.007***	.419	.195	.417	***.006***	.675	.109
# of felony arrests, age 27 (+)	effect	-14%	15%	33%	-1%	-17%	38%	-1%	19%	71%	36%	-27%	79%
	p-value	.813	.259	***.005***	.641	.724	***.006***	.642	.174	.178	***.011***	.781	***.058***
Jobless for more than 1 year, age 27 (-)	effect	33%	54%	24%	22%	4%	0%	23%	27%	47%	4%	8%	60%
	p-value	***.037***	***.011***	.170	.224	.453	.509	.234	.312	***.096***	.459	.345	.102
Ever tried drugs other than alcohol or weed, age 27 (-)	effect	-7%	-12%	32%	-9%	18%	40%	-17%	41%	37%	34%	-22%	49%
	p-value	.716	.690	***.060***	.733	.198	***.033***	.731	***.082***	.187	***.034***	.725	.125
# of misdemeanor violent crimes, age 40 (-)	effect	-3%	13%	34%	4%	-17%	35%	3%	21%	12%	36%	-5%	42%
	p-value	.580	.230	***.006***	.217	.787	***.008***	.372	.128	.392	***.004***	.650	***.082***
# of felony arrests, age 40 (-)	effect	2%	21%	33%	3%	-5%	34%	0%	29%	55%	32%	-18%	69%
	p-value	.435	***.090***	***.004***	.301	.569	***.006***	.508	***.073***	.117	***.005***	.759	***.017***
# of lifetime violent crimes, age 40 (-)	effect	-6%	13%	34%	2%	-18%	35%	1%	18%	26%	35%	-11%	50%
	p-value	.725	.199	***.005***	.372	.857	***.003***	.471	.174	.125	***.006***	.847	***.020***
Months in all marriages, age 40 (+)	effect	38%	35%	39%	17%	45%	34%	0%	79%	40%	32%	-1%	71%
	p-value	***.043***	***.012***	***.023***	.120	***.055***	***.030***	.469	***.008***	.112	***.030***	.500	***.007***

Notes: Percentages of the treatment effect explained by indices of traits are shown. No equality between coefficients of treatment and control groups is imposed. Decompositions are evaluated at the average level of the coefficients estimated for the treatment and control groups. One-sided *p*-values are based on 1000 bootstrap draws. *p*-values below 10% are in bold. Signs (+) and (-) represent the sign of the total treatment effect, which is shown in Table 1 of the text. Indices of Externalizing Behavior and Academic Motivation constructed in this paper are defined in Table 2 of the text. Indices of IQ and CAT are averages over IQ and CAT measures at ages 7, 8, and 9. [a]A part of the treatment effect explained by changes in the index. [b]A part of the treatment effect explained by changes in the coefficient. Signs "-" denote omitted estimates of models that make no sense since they have the early CAT on both right-hand and left-hand sides.

L Specification and Robustness Tests

This appendix presents supplementary analysis. Figure L.1 compares factor scores across genders, with p-values testing the equality of factor score means (denoted by p_m) and the equality of factor score distributions[22] (denoted by p_k) between males and females. The tests show that the factor scores have means and distributions that are comparable across genders, which suggests that both genders have similar skills whether they are in the treatment group or in the control group.[23]

Psychological measures are usually associated with substantial measurement error (e.g., Cunha and Heckman, 2008, and Cunha, Heckman and Schennach, 2010). Table L.1 demonstrates that, as expected, noise is generally high for the PBI measures. We calculate signal and noise for items used in model estimation in a similar fashion as Cunha, Heckman and Schennach (2010).

In the notation of this paper, we calculate signal as

$$\boldsymbol{S}^j_{M^j} = \frac{[\varphi^j_{m^j}]^2 \operatorname{Var}(\theta^j)}{[\varphi^j_{m^j}]^2 \operatorname{Var}(\theta^j) + \operatorname{Var}(\eta^j_{m^j})}, \tag{L-1}$$

and noise as

$$\boldsymbol{N}^j_{M^j} = \frac{\operatorname{Var}(\eta^j_{m^j})}{[\varphi^j_{m^j}]^2 \operatorname{Var}(\theta^j) + \operatorname{Var}(\eta^j_{m^j})}. \tag{L-2}$$

Tables L.2 and L.3 report specification tests for the outcome models. The tests show that the assumption that model coefficients are the same for treatment and control groups is empirically justified. We present Wald test statistics with p-values in parentheses for the specification tests. Refer to Section II.C for the discussion of motivation and identification related to these tests.

In the third column, we test whether treatment group factor loadings in equation (5) are the same as the control group factor loadings: $H_0 : \boldsymbol{\alpha}_{k,0} = \boldsymbol{\alpha}_{k,1}$, for each outcome $k \in K$,

[22] The p-values for the equality of distributions are obtained using the Kolmogorov-Smirnov test.

[23] Note that the Kolmogorov-Smirnov test results should be interpreted with caution, since for small samples the test is known to have low power.

where the subscript "0" denotes the control group, and "1" denotes the treatment group.

In the fourth column, we test whether the treatment group regression coefficients in equation (5) are the same as the control group coefficients: $H_0 : \boldsymbol{\beta}_{k,0} = \boldsymbol{\beta}_{k,1}$, for each $k \in K$, where the subscript "0" denotes the control group, and "1" denotes the treatment group.

Following the discussion of Section II.C, for the measurement equations we report tests of equality for intercepts and coefficients between treatment and control groups in Table L.4 (see equations (E-1) and (E-2)). Wald test statistics and the corresponding p-values are shown. For each factor $j \in \mathcal{J}_p$, we test whether the treatment and control groups have common intercepts in equation (E-2): $H_0 : \nu^j_{m^j,0} = \nu^j_{m^j,1}$, $\forall\ m^j \in \mathcal{M}^j / \{1\}$, where "0" denotes the control group and "1" denotes the treatment group. For each factor $j \in \mathcal{J}_p$, we also test whether the treatment and control groups have the same factor loadings: $H_0 : \varphi^j_{m^j,0} = \varphi^j_{m^j,1}$, $\forall\ m^j \in \mathcal{M}^j / \{1\}$, where "0" denotes the control group and "1" denotes the treatment group. Our results show that our assumptions of the equality of intercepts and coefficients are supported by the data.

Figure L.4 compares estimates based on three-step procedure used in this paper with estimates based on one-step maximum likelihood estimation. Tables L.5 and L.6 supplement Figure L.4. Results from both procedures are in close agreement, although p-values from the maximum likelihood procedure are generally lower. See Section IV.E for futher discussion.

Table L.7 presents factor loadings for the three-factor model using an alternative to quartimin called *geomin*.[24] It supplements Table H.2 in Appendix H. The estimates show that results of the exploratory factor analysis are robust to alternative methods of oblique rotation.

Table L.8 shows the factor loadings obtained through confirmatory factor analysis for the factor model described by Equation (8). The factor loadings are obtained via maximum likelihood estimation. The table shows that all loadings in the range 0.6–1.3, and statistically significant at the 1% level.

[24] Yates (1987*a*)

Table L.9 presents correlations among factors based on the MLE estimation of the measurement system. The table shows that for both males and females, there are statistically significant correlations between Cognition and Academic Motivation, and between Externalizing Behavior and Academic Motivation. However, the correlation between Cognition and Externalizing Behavior is not statistically significant.

Figure L.2 displays the quality of approximation of the decompositions demonstrated in Figures 6 and 7 of the paper. Tables L.10 and L.11 show the estimates of the decompositions,[25] while Figures 6 and 7 approximate the tables for a better visualization of the results by setting some statistically insignificant coefficients to zero. The components set to zero are the ones whose signs are opposite to those of the total treatment effects. We make this approximation because we cannot easily show negative terms of a sum in a simple bar graph, while those small and statistically insignificant terms that we equate to zero are not informative anyway. The histogram in Figure L.2 shows that our approximation is reasonable. The "quality of approximation," as defined in the notes to Figure L.2, ranges from 67% to 100%, with 3/4 of mass above 80% and with mean and median of 88%.

Tables L.12 and L.13 show the full set of estimates from the decompositions comparing the use of the California Achievement Test with that of Stanford-Binet IQ scores as a measure of intelligence in the measurement model. They supplement Figure L.3.

It is common in the literature to use achievement test scores rather than IQ scores as measures of cognition. Achievement Scores are highly loaded on personality skills (Borghans et al., 2011). We demonstrate how misleading the use of achievement scores can be by comparing decompositions using IQs with decompositions using CAT scores as measures of cognition. These two types of decompositions are substantially different. The achievement factor explains a much larger portion of the treatment effect than the factor that is based on IQ measures (see Figure L.3).[26] The result is not surprising. Indeed, CAT is loaded on

[25] Tables L.14 and L.15 show the corresponding attenuation-bias-corrected regression coefficients.

[26] Estimate based on the achievement test is numerically high, but still not statistically significant. We calculate these comparisons for a reduced sample size for which both IQ and CAT measures are non-missing.

personality skills likely including those that we cannot proxy. This makes the treatment effect on CAT higher ($E(\Delta\theta^{CAT}) > E(\Delta\theta^{IQ})$). Hence, it would be misleading to attribute stronger decompositions based on CAT to pure measures of cognition.

Figure L.5 presents the empirical CDFs of the factor scores. This figure supplements Figure 5 of the main paper, which shows the corresponding kernel density graphs. Refer to Section II.A of the main paper for a discussion of the treatment effect on the factor scores.

Tables L.10, L.11, L.14, and L.15 report the contributions of each of the improvements in Cognition, Externalizing Behavior, Academic Motivation, and other factors to the explanation of total treatment effects, as well as factor loadings and regression coefficients. These tables supplement Figures 6 and 7 of the main paper. Refer to section IV.C of the paper for discussion of contributions to the total treatment effect.

Tables L.16–L.19 test whether there are treatment effects on psychological traits. Table L.16 is devoted to cognition. The table shows statistically significant treatment effects on all measures of IQ for both genders at ages 4 and 5. At ages 6–10 we observe statistically significant effects on IQ only for females. Finally, at age 14, we observe statistically significant effect on the California Achievement Test for both genders.

Tables L.17–L.18 test for treatment effects based on PBI and YRS measures described in sections C and D. Two of them, namely Tables L.17 and L.18, show augmented measures that are averaged over ages 7–9 over non-missing values. We can see that, for females, a much larger set of measures is boosted than for males. Moreover, for males, boosted measures are primarily related to Externalizing Behavior (see Table L.17). The YRS measures show no effects for males and only a few effects for females, which makes YRS measures less likely candidates for expanding treatment effects of the program (see Table L.18).

Table L.19 shows treatment effects for various indices. By indicies we mean equally-weighted averages of trait measures as discussed in the text. We use the same measures to define alternative indices. First, we form PBI and YRS indicies as recommended by

the authors of PBI and YRS.[27] Then, we use an expert opinion documented in sections C and D to form indices approximating the Big Five personality traits. We form indices in two alternative ways but most results are robust to these differences.[28] Finally, we form two indices as defined in Table 2 of the main paper, Personal Behavior and Academic Motivation. Those indices approximate factors that are used for the main model of this paper.

Results in Table L.19 are in line with results of Tables L.17–L.18 and the rest of the paper. Among PBI and YRS original indices, only PBI Personal Behavior index shows a treatment effect for males, while a variety of indices show treatment effects for females. Similarly, we see many effects on Big Five traits for females, and virtually no effect for males.[29] For the indices representing factors constructed in this paper, for both genders we see statistically significant effect on Externalizing Behavior and a borderline significant effect on Academic Motivation. Figures L.6 and L.7 show the full set of decompositions of indices for all of the treatment effects.

[27] See Sections C and D for definitions of PBI and YRS indices.

[28] The first way is to use only measures dedicated to a particular trait to form an index for that trait. Under this approach, no measure is used twice for calculating indices, and all measures that are linked to more than one trait are unused. The second way is to use each measure that is linked to multiple traits in addition to dedicated measures. Under this approach, while dedicated measures are still used only once as before, measures that are linked to K traits are used K times to form K indices. For instance, if some measure is linked to both Conscientiousness and Agreeableness, it will be used to form two indices describing these traits. The advantage of the second methods is that more measures are used, which comes at a cost of less precise definition of a trait.

[29] We see an effect on Neuroticism for males with p-value of 0.079, but this effect is not robust to using dedicated measures only.

Table L.1: Measurement Errors of Items Used in the Factor Model

(Proportion Signal and Proportion Noise)

Item	Description	Age	Males		Females	
			Signal	Noise	Signal	Noise
Cognition						
Binet 7	Stanford-Binet intelligence scale	7	0.531	0.469	0.820	0.180
Binet 8	Stanford-Binet intelligence scale	8	0.694	0.306	0.776	0.224
Binet 9	Stanford-Binet intelligence scale	9	0.750	0.250	0.763	0.237
Externalizing Behavior						
PBI 27	Disrupts classroom procedures	7–9	0.745	0.255	0.789	0.211
PBI 28	Swears or uses obscene words	7–9	0.717	0.283	0.649	0.351
PBI 21	Steals	7–9	0.191	0.809	0.616	0.384
PBI 16	Lying or cheating	7–9	0.546	0.454	0.698	0.302
PBI 11	Influences others toward troublemaking	7–9	0.811	0.189	0.813	0.187
PBI 19	Aggressive toward peers	7–9	0.574	0.426	0.587	0.413
PBI 32	Teases and provokes students	7–9	0.639	0.361	0.382	0.618
Academic Motivation						
PBI 1	Shows initiative	7–9	0.784	0.216	0.751	0.249
PBI 4	Alert and interested in school work	7–9	0.957	0.043	0.991	0.009
PBI 25	Hesitant to try, or gives up easily	7–9	0.587	0.413	0.536	0.464
Sample			59		37	

Notes: Signal and noise are calculated based on formulas (L-2) and (L-1) in a similar fashion as in Flavio Cunha, James J. Heckman and Susanne M. Schennach (2010). "Age 7–9" stands for an average over non-missing observations at ages 7, 8, and 9.

Table L.2: Specification Tests, Males[(a)]

Outcome		H_0: $\alpha_1=\alpha_0$[(b)]	H_0: $\beta_1=\beta_0$[(c)]
CAT total, age 14*	test statistic	5.071	2.462
	p -value	(.289)	(.423)
# of misdemeanor arrests up to age 27	test statistic	.930	.617
	p -value	(.408)	(.524)
# of felony arrests up to age 27	test statistic	1.219	.195
	p -value	(.358)	(.821)
# of adult arrests up to age 27	test statistic	1.290	.101
	p -value	(.372)	(.890)
Monthly income at age 27	test statistic	12.017	2.109
	p -value	(.174)	(.489)
Use tobacco at age 27	test statistic	2.253	0.451
	p -value	(.214)	(.635)
# of misdemeanor arrests up to age 40	test statistic	1.819	.305
	p -value	(.293)	(.714)
# of felony arrests up to age 40	test statistic	.568	.581
	p -value	(.606)	(.569)
# of adult arrests up to age 40	test statistic	1.588	.140
	p -value	(.321)	(.879)
# of lifetime arrests	test statistic	1.426	.138
	p -value	(.352)	(.888)
Employed at age 40	test statistic	.411	5.681
	p -value	(.725)	(.162)

Notes: [(a)]Wald test statistics with p-values in parentheses for a number of specification tests. [(b)]Tests of whether the treatment group factor loadings in Equation (6) are the same as the control group factor loadings: $H_0 : \boldsymbol{\alpha}_0 = \boldsymbol{\alpha}_1$, for each outcome Y, where the subscript "0" denotes the control group, and "1" denotes the treatment group. [(c)]Tests of whether the treatment group regression coefficients in Equation (6) are the same as the control group coefficients: $H_0 : \boldsymbol{\beta}_0 = \boldsymbol{\beta}_1$, for each outcome, where the subscript "0" denotes the control group, and "1" denotes the treatment group.

Table L.3: Specification Tests, Females[(a)]

Outcome		H_0: $\alpha_1=\alpha_0$[(b)]	H_0: $\beta_1=\beta_0$[(c)]
CAT total at age 8	test statistic	1.011	1.092
	p -value	(.545)	(.589)
CAT total at age 14	test statistic	5.636	3.671
	p -value	(.419)	(.593)
Any special education up to age 14	test statistic	.306	2.053
	p -value	(.796)	(.671)
Mentally impaired at least once up to age 19	test statistic	.737	5.579
	p -value	(.637)	(.563)
# of misdemeanor violent crimes up to age 27	test statistic	1.270	.696
	p -value	(.660)	(.616)
# of felony arrests up to age 27	test statistic	1.157	.290
	p -value	(.408)	(.662)
Jobless for more than 1 year up to age 27	test statistic	2.701	.763
	p -value	(.429)	(.707)
Ever tried drugs other than alcohol or weed up to age 27	test statistic	.961	.649
	p -value	(.554)	(.702)
# of misd. violent crimes up to age 40	test statistic	1.451	.976
	p -value	(.563)	(.533)
# of felony arrest up to age 40	test statistic	.788	.887
	p -value	(.572)	(.495)
# of lifetime violent crimes up to age 40	test statistic	2.393	.793
	p -value	(.546)	(.594)
Months in all marriages up to age 40	test statistic	.634	1.104
	p -value	(.652)	(.601)

Notes: [(a)]Wald test statistics with p-values in parentheses for a number of specification tests. [(b)]Tests of whether the treatment group factor loadings in Equation (6) are the same as the control group factor loadings: $H_0 : \boldsymbol{\alpha}_0 = \boldsymbol{\alpha}_1$, for each outcome Y, where the subscript "0" denotes the control group, and "1" denotes the treatment group. [(c)]Tests of whether the treatment group regression coefficients in Equation (6) are the same as the control group coefficients: $H_0 : \boldsymbol{\beta}_0 = \boldsymbol{\beta}_1$, for each outcome, where the subscript "0" denotes the control group, and "1" denotes the treatment group.

Table L.4: Testing the Equality of Intercepts and Coefficients for Treatment and Control Groups in the Measurement Equations[(a)]

			Intercepts[(b)]		Coefficients[(c)]	
Factor	Age		Males	Females	Males	Females
Cognition	7–9	test statistic	3.057	.126	.857	.672
		p- value	(.217)	(.939)	(.676)	(.715)
Externalizing Behavior	7–9	test statistic	10.620	2.350	7.705	6.001
		p- value	(.101)	(.885)	(.261)	(.423)
Academic Motivation	7–9	test statistic	2.354	2.911	.413	1.231
		p- value	(.308)	(.233)	(.814)	(.540)

Notes: [(a)]Wald test statistics and the corresponding p-values are shown. [(b)]For each factor $j \in \mathcal{J}_p$, we test whether treatment and control groups have common intercepts in Equation (E-2): $H_0 : \nu^j_{m^j,0} = \nu^j_{m^j,1}$, $\forall$ $m^j \in \mathcal{M}^j \setminus \{1\}$, where "0" denotes the control group and "1" denotes the treatment group. [(c)]For each factor $j \in \mathcal{J}_p$, we test whether treatment and control groups have the same coefficients in Equation (E-2): $H_0 : \varphi^j_{m^j,0} = \varphi^j_{m^j,1}$, $\forall$ $m^j \in \mathcal{M}^j \setminus \{1\}$, where "0" denotes the control group and "1" denotes the treatment group.

Table L.5: Decompositions of Treatment Effects, Factor Scores Versus MLE, Males

Outcome		Statistic	Cognition effect[a]	Externalizing Behavior[b]	Academic Motivation[c]	Total explained effect of non-cognitive skills[d]	Total explained effect[e]	Residual effect[f]	Available Observations
# of misdemeanor arrests, age 27	SCORE	effect	.023	-.447 *	-.031	-.478 *	-.455	-1.161	59
		p- value	.567	***.071***	.557	***.084***	.115	.114	
	MLE	effect	.017	-.418 *	-.056	-.475 *	-.458 *	-1.126 *	59
		p- value	.411	***.091***	.342	***.057***	***.072***	***.072***	
# of felony arrests, age 27	SCORE	effect	.055	-.486 *	.045	-.441 *	-.385	-.612	59
		p- value	.603	***.071***	.654	***.098***	.142	.246	
	MLE	effect	.048	-.451 *	.011	-.440 *	-.392	-.591	59
		p- value	.333	***.085***	.467	***.077***	.119	.223	
# of misdemeanor arrests, age 40	SCORE	effect	.056	-.883	.040	-.843	-.787	-2.883 *	59
		p- value	.553	.136	.454	.148	.180	***.088***	
	MLE	effect	.037	-.810	-.021	-.831	-.794	-3.082 *	59
		p- value	.425	.132	.475	.105	.131	***.067***	
# of felony violent crimes, age 40	SCORE	effect	.056	-.640 *	.060	-.579 *	-.523	-.305	59
		p- value	.575	***.056***	.643	***.082***	.122	.403	
	MLE	effect	.045	-.597 *	.018	-.580 *	-.535	-.302	59
		p- value	.353	***.079***	.456	***.074***	.104	.378	

Notes: "SCORE" denotes a three-step estimation method using factor scores as described in the main paper. "MLE" denotes a one-step maximum likelihood estimation method where both measurement system and outcome equation are estimated simultaneously. Estimated are the following population components of the models: (a) $\alpha_k^C E(\theta^C(1) - \theta^C(0))$; (b) $\alpha_k^E E(\theta^E(1) - \theta^E(0))$; (c) $\alpha_k^I E(\theta^A(1) - \theta^A(0))$; (d) $\alpha_k^E E(\theta^E(1) - \theta^E(0)) + \alpha_k^A E(\theta^A(1) - \theta^A(0))$; (e) $\boldsymbol{\alpha}_k E(\boldsymbol{\theta}(1) - \boldsymbol{\theta}(0))$; (f) τ_k, where "C" stands for "Cognition", "E" stands for "Externalizing Behavior", "A" stands for "Academic Motivation". One-sided bootstrap p-values are reported. p-values below 0.1 are in bold italics. The number of bootstrap iterations is 1000. Stars denote significance levels: ** - 5% and * - 10%. $\boldsymbol{X}$ includes three background variables at entry: mother's employment, father's presence in the household, and family's SES. "CAT total" denotes the California Achievement Test total score.

Table L.6: Decompositions of Treatment Effects, Factor Scores Versus MLE, Females

Outcome		Statistic	Cognition effect[(a)]	Externalizing Behavior[(b)]	Academic Motivation[(c)]	Total explained effect of non-cognitive skills[(d)]	Total explained effect[(e)]	Residual effect[(f)]	Available Observations
# if misdemeanor violent crimes, age 27	SCORE	effect	.050	-.572 *	.059	-.513	-.463	-.353	37
		p- value	.447	***.099***	.840	.117	.137	.305	
	MLE	effect	.048	-.546 **	.038	-.509 **	-.461 **	-.441	37
		p- value	.331	***.020***	.395	***.027***	***.041***	.129	
# of felony arrests, age 27	SCORE	effect	.013	-.239	.055	-.183	-.171	-.172	37
		p- value	.493	.120	.907	.125	.160	.319	
	MLE	effect	.019	-.234 **	.044	-.190 *	-.171	-.182	37
		p- value	.385	***.048***	.308	***.088***	.111	.231	
# if misdemeanor violent crimes, age 40	SCORE	effect	.050	-.571 *	.032	-.539 *	-.489 *	-.289	37
		p- value	.437	***.066***	.787	***.079***	***.093***	.371	
	MLE	effect	.045	-.550 **	.017	-.533 **	-.488 **	-.337	37
		p- value	.340	***.020***	.451	***.022***	***.033***	.196	
# of felory arrests, age 40	SCORE	effect	.028	-.312 **	.031	-.281 *	-.253 *	-.177	37
		p- value	.437	***.050***	.361	***.065***	***.059***	.369	
	MLE	effect	.031	-.311 **	.025	-.285 **	-.254 *	-.179	37
		p- value	.321	***.025***	.382	***.034***	***.052***	.230	

Notes: "FACTOR" denotes a three-step estimation method using factor scores as described in the main paper. "MLE" denotes a one-step maximum likelihood estimation method where both measurement system and outcome equation are estimated simultaneously. Estimated are the following population components of the models: (a) $\alpha_k^C E(\theta^C(1) - \theta^C(0))$; (b) $\alpha_k^E E(\theta^E(1) - \theta^E(0))$; (c) $\alpha_k^I E(\theta^A(1) - \theta^A(0))$; (d) $\alpha_k^E E(\theta^E(1) - \theta^E(0)) + \alpha_k^A E(\theta^A(1) - \theta^A(0))$; (e) $\boldsymbol{\alpha}_k E(\boldsymbol{\theta}(1) - \boldsymbol{\theta}(0))$; (f) τ_k, where "C" stands for "Cognition", "E" stands for "Externalizing Behavior", "A" stands for "Academic Motivation". One-sided bootstrap p-values are reported. p-values below 0.1 are in bold italics. The number of bootstrap iterations is 1000. Stars denote significance levels: ** - 5% and * - 10%. $\boldsymbol{X}$ includes three background variables at entry: mother's employment, father's presence in the household, and family's SES. "CAT total" denotes the California Achievement Test total score.

Table L.7: Factor Loadings of a Three-Factor Model After Geomin Rotation

	Males						Females						Pooled					
	Cognition	Standard Error	Externalizing Behavior	Standard Error	Academic Motivation	Standard Error	Cognition	Standard Error	Externalizing Behavior	Standard Error	Academic Motivation	Standard Error	Cognition	Standard Error	Externalizing Behavior	Standard Error	Academic Motivation	Standard Error
Cognition																		
Stanford Binet, age 7	**-.661**	*(.106)*	-.016	*(.062)*	-.121	*(.131)*	**.890**	*(.134)*	-.114	*(.096)*	.115	*(.163)*	**.777**	*(.068)*	-.040	*(.042)*	.098	*(.089)*
Stanford Binet, age 8	**-.686**	*(.093)*	-.089	*(.092)*	-.217	*(.120)*	**.853**	*(.068)*	.008	*(.063)*	.041	*(.088)*	**.798**	*(.064)*	-.037	*(.037)*	.114	*(.087)*
Stanford Binet, age 9	**-.932**	*(.058)*	.087	*(.090)*	-.011	*(.033)*	**.886**	*(.075)*	.074	*(.098)*	-.067	*(.089)*	**.900**	*(.038)*	.083	*(.062)*	-.027	*(.022)*
Externalizing Behavior																		
Disrupts classroom procedures	.023	*(.051)*	**.759**	*(.071)*	-.231	*(.107)*	-.101	*(.105)*	**.830**	*(.089)*	.157	*(.140)*	-.043	*(.052)*	**.759**	*(.060)*	.248	*(.094)*
Swears or uses obscene words	.099	*(.083)*	**.708**	*(.078)*	-.287	*(.107)*	.021	*(.098)*	**.699**	*(.108)*	.203	*(.165)*	-.045	*(.060)*	**.683**	*(.069)*	.284	*(.100)*
Steals	-.018	*(.134)*	**.360**	*(.133)*	-.145	*(.151)*	-.014	*(.105)*	**.743**	*(.104)*	.122	*(.160)*	.057	*(.103)*	**.461**	*(.096)*	.151	*(.122)*
Lying or cheating	.117	*(.103)*	**.548**	*(.101)*	-.375	*(.118)*	-.052	*(.105)*	**.790**	*(.097)*	.144	*(.152)*	-.062	*(.072)*	**.597**	*(.077)*	.340	*(.101)*
Influences others toward troublemaking	-.043	*(.067)*	**.902**	*(.046)*	-.033	*(.077)*	-.039	*(.096)*	**.924**	*(.048)*	-.021	*(.057)*	.018	*(.050)*	**.895**	*(.038)*	.050	*(.069)*
Aggressive toward peers	-.336	*(.099)*	**.826**	*(.072)*	.092	*(.084)*	.075	*(.090)*	**.931**	*(.082)*	-.344	*(.150)*	.225	*(.081)*	**.817**	*(.053)*	-.110	*(.076)*
Teases or provokes students	-.126	*(.103)*	**.814**	*(.060)*	.005	*(.064)*	.062	*(.132)*	**.718**	*(.114)*	-.220	*(.181)*	.079	*(.085)*	**.756**	*(.056)*	-.049	*(.063)*
Academic Motivation																		
Shows Initiative	-.043	*(.057)*	-.070	*(.058)*	**-.918**	*(.051)*	.087	*(.214)*	-.053	*(.065)*	**.898**	*(.116)*	.014	*(.029)*	-.106	*(.058)*	**.937**	*(.039)*
Alert and interested in school work	-.061	*(.059)*	.061	*(.056)*	**-.912**	*(.047)*	.239	*(.229)*	.117	*(.091)*	**.769**	*(.157)*	.086	*(.058)*	.053	*(.037)*	**.899**	*(.040)*
Hesitant to try, or gives up easily	-.046	*(.080)*	.185	*(.108)*	**-.686**	*(.087)*	.300	*(.212)*	.057	*(.102)*	**.547**	*(.167)*	.116	*(.086)*	.139	*(.084)*	**.663**	*(.077)*
Sample size			59						37						96			

Notes: Factor loadings based on the exploratory factor analysis with geomin rotation (Yates, 1987*b*) are shown. Maximum likelihood asymptotic standard errors are in parentheses. Factor loadings relating factors to corresponding potential dedicated measures are in bold. See Table H.2 for a similar result based on direct quartimin oblique rotation.

Table L.8: Estimates of Factor Loadings for the Measurement System

	Males		Females	
	Coefficient	Standard Error	Coefficient	Standard Error
Cognition				
Stanford Binet, age 7	1	-	1	-
Stanford Binet, age 8	1.283 ***	(.224)	.932 ***	(.128)
Stanford Binet, age 9	1.154 ***	(.189)	.698 ***	(.096)
Externalizing Behavior				
Disrupts classroom procedures	1	-	1	-
Swears or uses obscene words	1.051 ***	(.124)	1.042 ***	(.159)
Steals	.565 ***	(.165)	1.062 ***	(.178)
Lying or cheating	.906 ***	(.135)	1.080 ***	(.159)
Influences others toward troublemaking	1.162 ***	(.122)	1.142 ***	(.139)
Aggressive toward peers	.974 ***	(.138)	.922 ***	(.161)
Teases or provokes students	.961 ***	(.125)	.990 ***	(.235)
Academic Motivation				
Shows Initiative	1	-	1	-
Alert and interested in school work	1.121 ***	(.096)	1.115 ***	(.124)
Hesitant to try, or gives up easily	.909 ***	(.120)	.857 ***	(.156)
Tucker-Lewis index (TLI)[(b)]	.991		.975	
Comparative fit index (CFI)[(c)]	.837		.727	
Standardized root-mean-square-residual (RMSR)[(d)]	.085		.091	
Root mean square error of approximation (RMSEA)[(e)]	.071		.125	
Sample size	59		37	

Notes: [(a)]We present maximum likelihood estimates of the measurement system (described by Equation (8)). Standard errors are shown in parentheses. [(b)]TLI (Tucker and Lewis, 1973) ranges from zero to one, with values close to one indicating good fit. [(c)]Like TLI, the CFI (Bentler, 1990*a*,*b*) ranges from zero to one, with values close to one showing good fit. [(d)]RMSR (Jöreskog and Sörbom, 1986) ranges from zero to one, with values close to zero showing good fit. [(e)]Like RMSR, the RMSEA (Browne and Cudeck, 1992; Steiger, 1990) ranges for zero to one, with values close to zero showing good fit.

Table L.9: Correlations among Factors

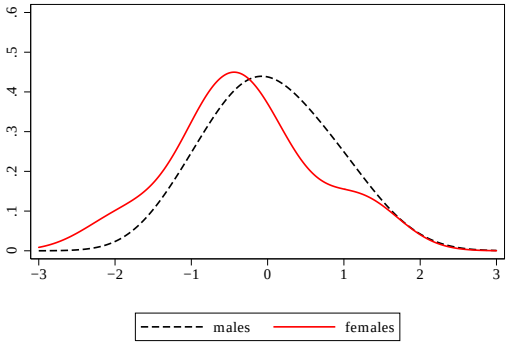
		Males			Females		
		Cognition	Externalizing Behavior	Academic Motivation	Cognition	Externalizing Behavior	Academic Motivation
Cognition	coefficient	1			1		
	std. error	*(–)*			*(–)*		
Externalizing Behavior	coefficient	.099	1		.254	1	
	std. error	*(.144)*	*(–)*		*(.168)*	*(–)*	
Academic Motivation	coefficient	.509 ***	.536 ***	1	.651 ***	.516 ***	1
	std. error	*(.110)*	*(.101)*	*(–)*	*(.105)*	*(.127)*	*(–)*
Sample			59			37	

Notes: Correlations are shown with standard errors reported in parentheses. Stars denote statistical significance of the correlation: "***", 1% level; "**", 5% level; "*", 10% level.

Figure L.1: Gender Comparisons of Factor Scores

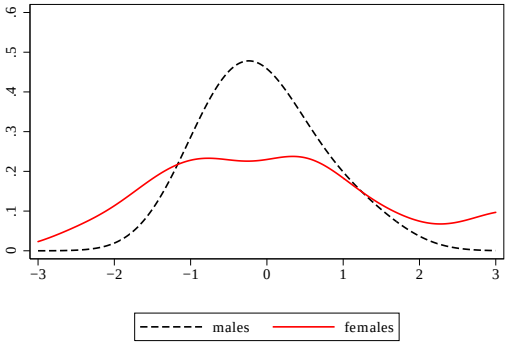
(a) Cognition, Control Group

$p_m = .190; p_k = .312$



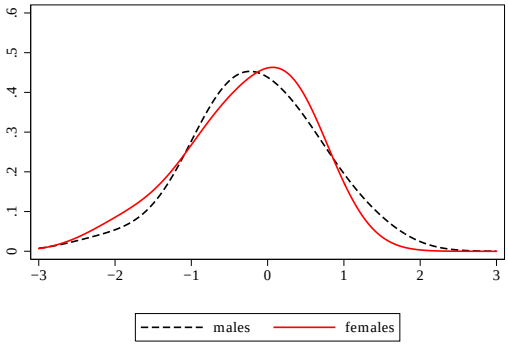
(b) Cognition, Treatment Group

$p_m = .363; p_k = .596$



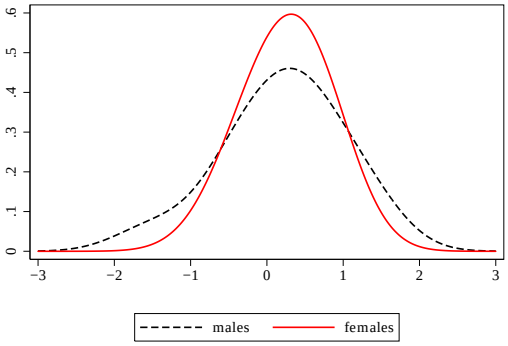
(c) Externalizing Behavior, Control Group

$p_m = .555; p_k = .965$



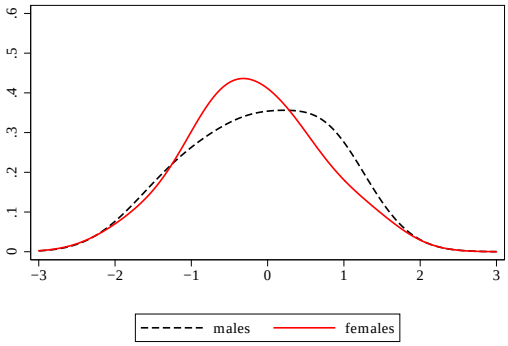
(d) Externalizing Behavior, Treatment Group

$p_m = .693; p_k = .589$



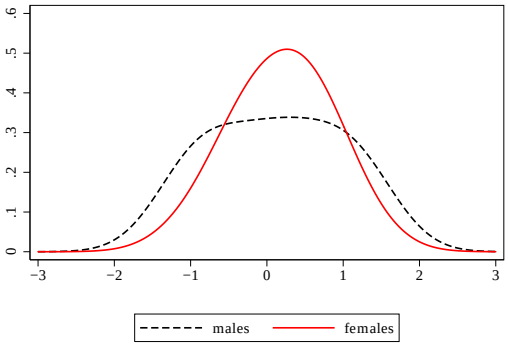
(e) Academic Motivation, Control Group

$p_m = .596; p_k = .532$



(f) Academic Motivation, Treatment Group

$p_m = .668; p_k = .458$



Notes: Kernel density functions of Bartlett (1937) factor scores are shown. (See the discussion in Web Appendix F.) p_m is the p-value testing the hypothesis that factor scores have equal means across gender. p_k is the p-value testing the hypothesis that factor scores have equal distributions across gender. Higher personality scores correspond to more socially desirable behaviors like less aggression or more interest in schooling. Measures of factors are normalized for a pooled sample of males and females to capture gender differences.

Table L.10: Decompositions of Treatment Effects on Outcomes, Males

Outcome	Statistic	Cognition effect[(a)]	Externalizing Behavior[(b)]	Academic Motivation[(c)]	Total explained effect of non-cognitive skills[(d)]	Total explained effect[(e)]	Residual effect[(f)]	Available Observations
CAT total at age 14, end of grade 8 (+)	effect	-.073	-.074	.144	.070	-.004	.728 **	45
	p-value	.687	.690	.161	.268	.480	***.013***	
# of misdemeanor arrests, age 27 (-)	effect	.023	-.447 *	-.031	-.478 *	-.455	-1.161	59
	p-value	.567	***.071***	.557	***.084***	.115	.114	
# of felony arrests, age 27 (-)	effect	.055	-.486 *	.045	-.441 *	-.385	-.612	59
	p-value	.603	***.071***	.654	***.098***	.142	.246	
# of adult arrests (misd.+fel.), age 27 (-)	effect	.079	-.932 *	.014	-.919 *	-.840	-1.774	59
	p-value	.594	***.062***	.474	***.074***	.108	.144	
Monthly income, age 27 (+)	effect	-.087	.254 *	-.053	.202	.115	1.110 **	55
	p-value	.690	***.089***	.730	.144	.334	***.027***	
Use tobacco, age 27 (-)	effect	.016	-.121 **	.033	-.088	-.072	-.161	57
	p-value	.643	***.046***	.628	.100	.180	.141	
# of misdemeanor arrests, age 40 (-)	effect	.056	-.883	.040	-.843	-.787	-2.883 *	59
	p-value	.553	.136	.454	.148	.180	***.088***	
# of felony arrests, age 40 (-)	effect	.056	-.640 *	.060	-.579 *	-.523	-.305	59
	p-value	.575	***.056***	.643	***.082***	.122	.403	
# of adult arrests (misd.+fel.), age 40 (-)	effect	.112	-1.523 *	.101	-1.422	-1.310	-3.188	59
	p-value	.556	***.086***	.479	.108	.142	.149	
# of lifetime arrests, age 40 (-)	effect	.099	-1.727 *	.049	-1.678 *	-1.579	-2.831	59
	p-value	.543	***.077***	.597	***.099***	.121	.204	
Employed, age 40 (+)	effect	-.025	.084 *	-.056	.028	.003	.336 **	54
	p-value	.667	***.085***	.834	.353	.454	***.018***	

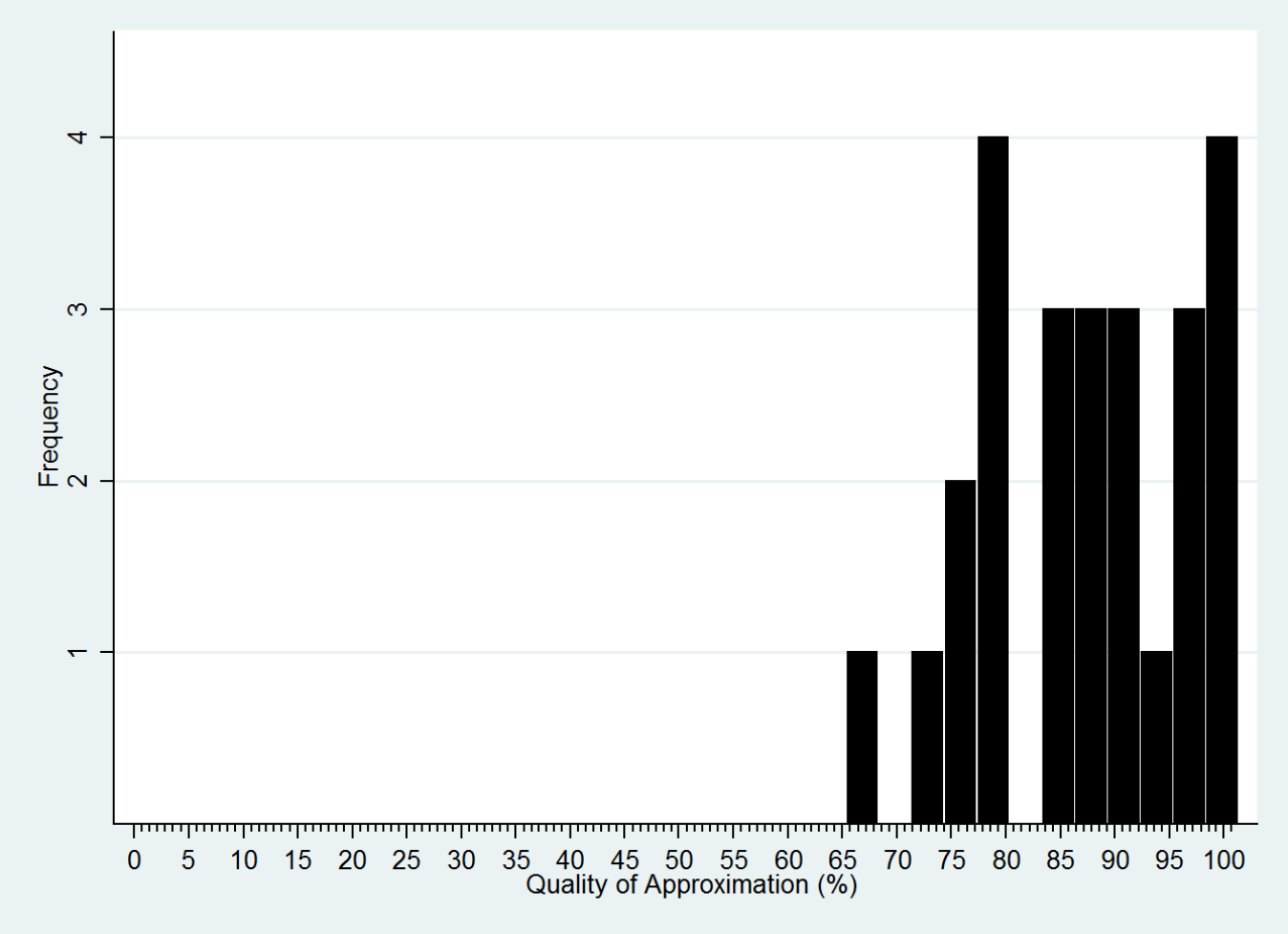
Notes: Estimated are the following population components of the models: (a) $\alpha_k^C E(\theta^C(1) - \theta^C(0))$; (b) $\alpha_k^E E(\theta^E(1) - \theta^E(0))$; (c) $\alpha_k^I E(\theta^A(1) - \theta^A(0))$; (d) $\alpha_k^E E(\theta^E(1) - \theta^E(0)) + \alpha_k^A E(\theta^A(1) - \theta^A(0))$; (e) $\boldsymbol{\alpha}_k E(\boldsymbol{\theta}(1) - \boldsymbol{\theta}(0))$; (f) τ_k, where "C" stands for "Cognition", "E" stands for "Externalizing Behavior", "A" stands for "Academic Motivation". One-sided bootstrap p-values are reported. p-values below 0.1 are in bold italics. The number of bootstrap iterations is 1000. Stars denote significance levels: ** - 5% and * - 10%. $\boldsymbol{X}$ includes three background variables at entry: mother's employment, father's presence in the household, and family's SES. (*) "CAT total" denotes the California Achievement Test total score.

Table L.11: Decompositions of Treatment Effects on Outcomes, Females

Outcome	Statistic	Cognition effect[(a)]	Externalizing Behavior[(b)]	Academic Motivation[(c)]	Total explained effect of non-cognitive skills[(d)]	Total explained effect[(e)]	Residual effect[(f)]	Available Observations
CAT total, age 8 (+)	effect	.131	-.071	.271 *	.200	.332	.498	35
	p-value	.153	.450	***.057***	.164	.127	.283	
CAT total, age 14 (+)	effect	.092	-.237	.354	.117	.209	.929	31
	p-value	.256	.533	.528	.226	.204	.232	
Any special education, age 14 (-)	effect	-.024	.063	-.082	-.019	-.044	-.463 *	37
	p-value	.344	.559	.533	.379	.320	***.071***	
Mentally impaired at least once, age 19 (+)	effect	-.024	.120	-.121 **	-.001	-.024	-.323	33
	p-value	.339	.681	***.042***	.489	.394	.109	
# of misdemeanor violent crimes, age 27 (-)	effect	.050	-.572 *	.059	-.513	-.463	-.353	37
	p-value	.447	***.099***	.840	.117	.137	.305	
# of felony arrests, age 27 (+)	effect	.013	-.239	.055	-.183	-.171	-.172	37
	p-value	.493	.120	.907	.125	.160	.319	
Jobless for more than 1 year, age 27 (-)	effect	.084	.025	-.183	-.158	-.074	-.316	36
	p-value	.620	.373	.497	.127	.214	.157	
Ever tried drugs other than alcohol or weed, age 27 (-)	effect	-.025	-.077	.048	-.029	-.055	-.153	34
	p-value	.199	.228	.884	.309	.228	.150	
# of misdemeanor violent crimes, age 40 (-)	effect	.050	-.571 *	.032	-.539 *	-.489 *	-.289	37
	p-value	.437	***.066***	.787	***.079***	***.093***	.371	
# of felony arrests, age 40 (-)	effect	.028	-.312 **	.031	-.281 *	-.253 *	-.177	37
	p-value	.437	***.050***	.361	***.065***	***.059***	.369	
# of lifetime violent crimes, age 40 (-)	effect	.058	-.646 **	.065	-.581 *	-.524 *	-.342	37
	p-value	.532	***.046***	.843	***.062***	***.075***	.320	
Months in all marriages, age 40 (+)	effect	13.040	7.197	4.117	11.315	24.354	38.167	36
	p-value	.185	.224	.269	.203	.134	.352	

Notes: One-sided bootstrap p-values are reported. p-values below 0.1 are in bold italics. The number of bootstrap iterations is 1000. Stars denote significance levels: ** - 5% and * - 10%. "CAT total" denotes the California Achievement Test total score.

Figure L.2: Quality of the Approximation Associated with the Decomposition Figures



Notes: The chart graphically represents the degree of approximation used for the decompositions presented in Figures 6 and 7 of the main paper. Tables L.10–L.15 show the actual estimates, while Figures 6 and 7 approximate the estimates by setting some small and statistically insignificant components to zero. Estimates are set to zero if they have the opposite sign of the total treatment effect. We make this approximation because we cannot easily show negative terms of a sum in a simple bar graph. The histogram shows the frequencies of "qualities" of these approximations for the set of outcomes Y_k, $k \in \{1, \cdots, K\}$ used in this paper. Qualities for males and females are pooled together in this histogram. "Quality" is defined as $(1 - \frac{|\omega_k|}{|\Delta_k|\boldsymbol{X}|}) \cdot 100\%$, where $|\Delta_k|\boldsymbol{X}|$ is the absolute value of the total conditional treatment effect on outcome Y_k; ω_k is the effect $E(\theta_1^j - \theta_0^j)$ of one individual skill that we set to zero in our approximation, where j is some element of $\mathcal{J}$. If no term is set to zero for that outcome Y_k, then $\omega_k = 0$, and the goodness is 100%.

Table L.12: Decompositions of Treatment Effects by Achievement and IQ, Males

Outcome		Statistic	Cognition effect[(a)]	Externalizing Behavior[(b)]	Academic Motivation[(c)]	Total explained effect of non-cognitive skills[(d)]	Total explained effect[(e)]	Residual effect[(f)]	Available Observations
CAT total, age 14	IQ	effect	-.036	-.085	.264 *	.178	.142	.722 **	41
		p-value	.535	.608	***.051***	.131	.267	***.024***	
	CAT	effect	.216	-.026	.093	.067	.284	.516 *	41
		p-value	.132	.507	.242	.283	.102	***.070***	
# of misdemeanor arrests, age 27	IQ	effect	.007	-.817 **	-.111	-.928 **	-.921 **	-.247	52
		p-value	.537	***.021***	.502	***.020***	***.033***	.403	
	CAT	effect	-.278	-.860 **	.171	-.689	-.968 **	-.214	52
		p-value	.203	***.023***	.678	.145	***.039***	.418	
# of misdemeanor arrests, age 40	IQ	effect	.028	-1.275	.083	-1.192	-1.163	-2.248	52
		p-value	.528	.120	.623	.125	.144	.188	
	CAT	effect	-.523	-1.303	.541	-.763	-1.285	-2.131	52
		p-value	.288	.118	.687	.297	.140	.195	
# of lifetime arrests, age 40	IQ	effect	.033	-2.809 **	.009	-2.800 **	-2.767 *	-1.287	52
		p-value	.494	***.044***	.605	***.044***	***.054***	.374	
	CAT	effect	-.647	-2.848 **	.578	-2.270	-2.918 *	-1.153	52
		p-value	.304	***.050***	.659	.169	***.061***	.373	

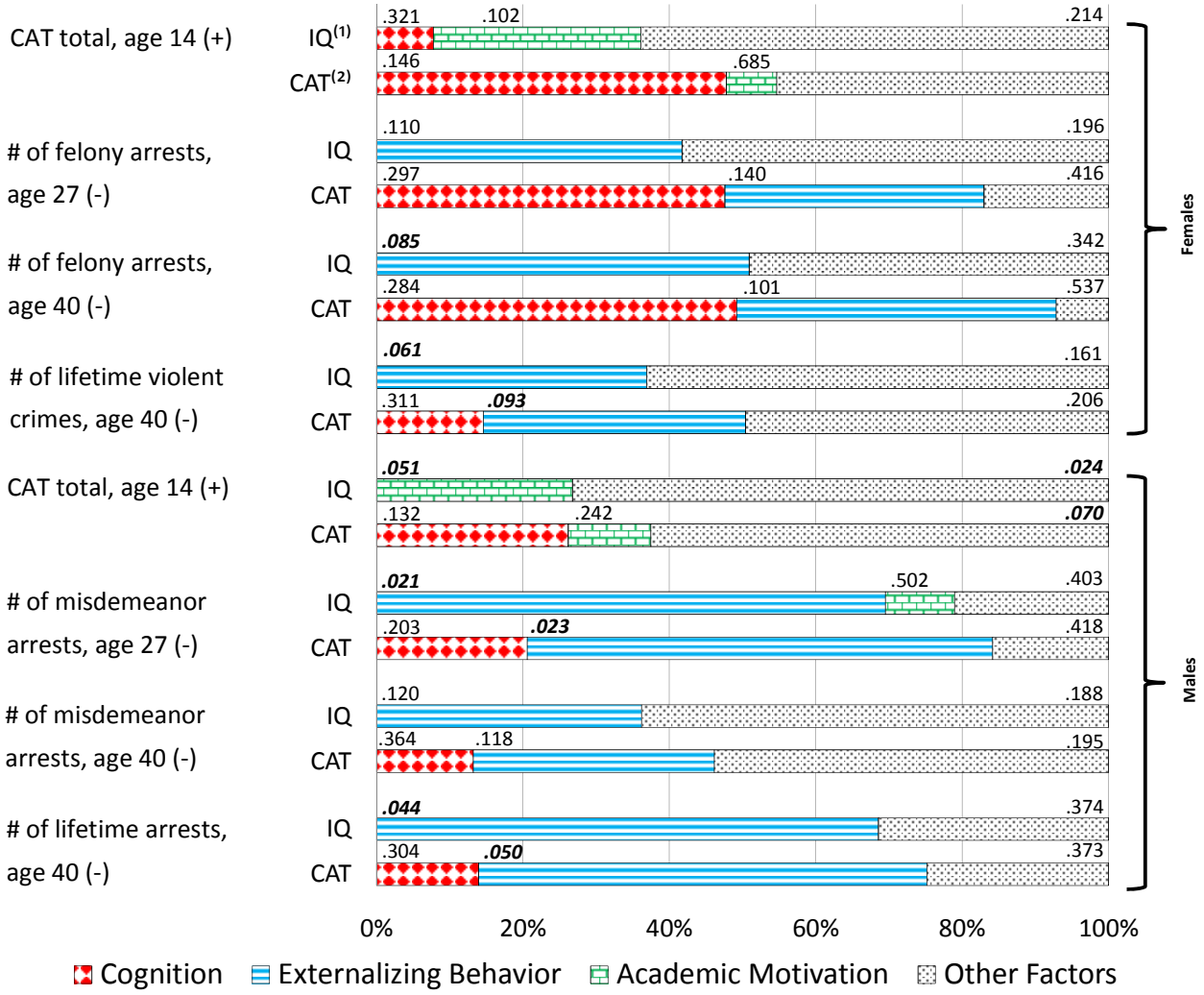
Notes: "IQ" denotes a model where cognition is measured by Stanford-Binet IQ. "CAT" denotes a model where cognition is measured by the California Achievement Test. In both cases, measures at ages 7, 8, and 9 are used. Estimated are the following population components of the models: (a) $\alpha_k^C E(\theta^C(1) - \theta^C(0))$; (b) $\alpha_k^E E(\theta^E(1) - \theta^E(0))$; (c) $\alpha_k^I E(\theta^A(1) - \theta^A(0))$; (d) $\alpha_k^E E(\theta^E(1) - \theta^E(0)) + \alpha_k^A E(\theta^A(1) - \theta^A(0))$; (e) $\boldsymbol{\alpha}_k E(\boldsymbol{\theta}(1) - \boldsymbol{\theta}(0))$; (f) τ_k, where "C" stands for "Cognition", "E" stands for "Externalizing Behavior", "A" stands for "Academic Motivation". One-sided bootstrap p-values are reported. p-values below 0.1 are in bold italics. The number of bootstrap iterations is 1000. Stars denote significance levels: ** - 5% and * - 10%. $\boldsymbol{X}$ includes three background variables at entry: mother's employment, father's presence in the household, and family's SES. "CAT total" denotes the California Achievement Test total score.

Table L.13: Decompositions of Treatment Effects by Achievement and IQ, Females

Outcome		Statistic	Cognition effect[(a)]	Externalizing Behavior[(b)]	Academic Motivation[(c)]	Total explained effect of non-cognitive skills[(d)]	Total explained effect[(e)]	Residual effect[(f)]	Available Observations
CAT total, age 14	IQ	effect	.090	-.176	.328	.152	.243	.740	24
		p-value	.321	.544	.102	.285	.271	.214	
	CAT	effect	.525	-.086	.076	-.010	.515	.498	24
		p-value	.146	.522	.685	.428	.169	.399	
# of felony arrests, age 27	IQ	effect	.009	-.221	.033	-.188	-.179	-.309	30
		p-value	.454	.110	.845	.134	.147	.196	
	CAT	effect	-.343	-.255	.225	-.030	-.373	-.123	30
		p-value	.297	.140	.828	.311	.179	.416	
# of felony arrests, age 40	IQ	effect	.031	-.237 *	.011	-.226	-.194	-.229	30
		p-value	.514	***.085***	.295	.114	.136	.342	
	CAT	effect	-.307	-.273	.198	-.074	-.382	-.045	30
		p-value	.284	.101	.395	.316	.144	.537	
# of lifetime violent crimes, age 40	IQ	effect	.011	-.236 *	.108	-.128	-.117	-.403	30
		p-value	.547	***.061***	.865	.167	.202	.161	
	CAT	effect	-.102	-.249 *	.172	-.077	-.179	-.346	30
		p-value	.311	***.093***	.424	.292	.254	.206	

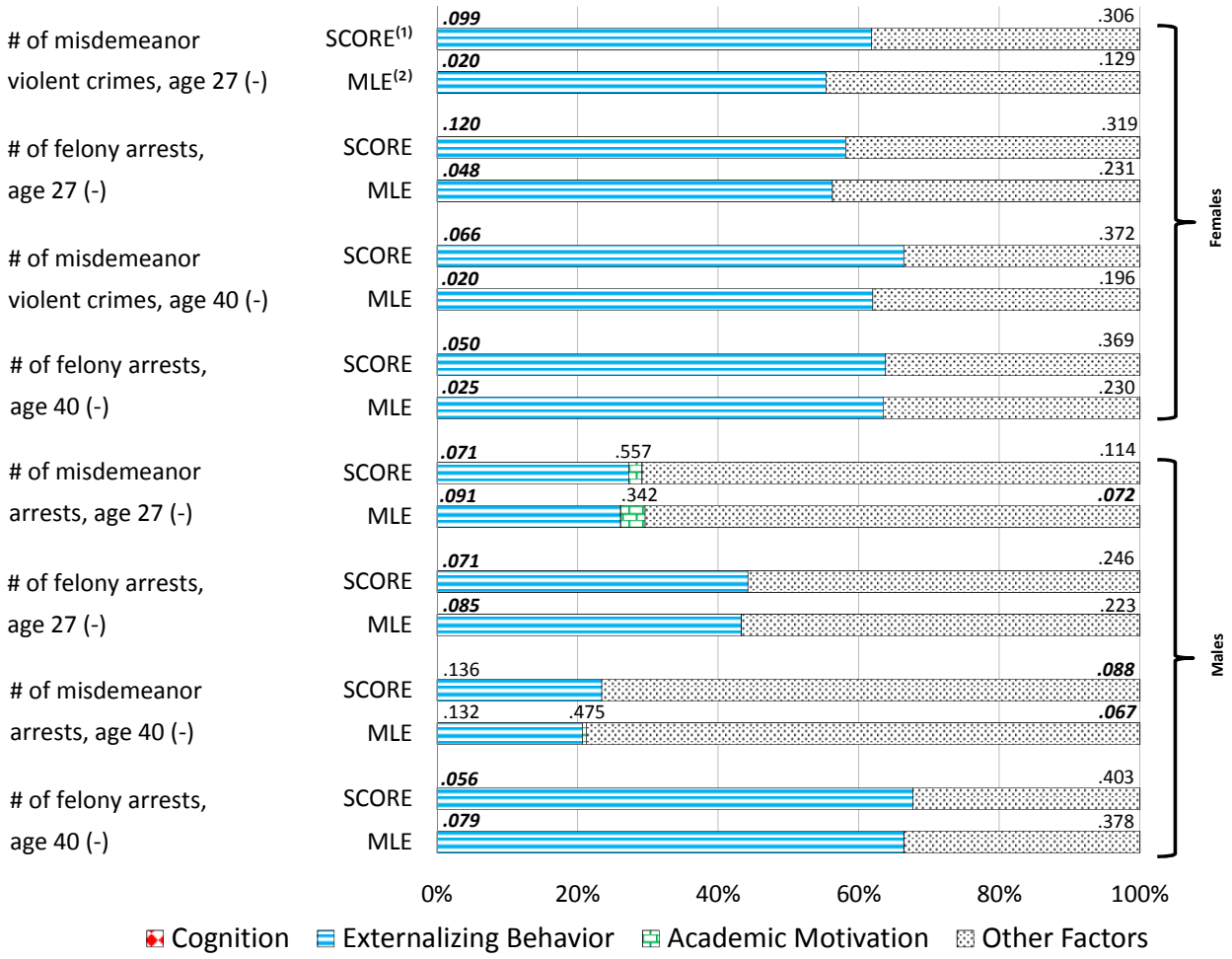
Notes: "IQ" denotes a model where cognition is measured by Stanford-Binet IQ. "CAT" denotes a model where cognition is measured by the California Achievement Test. In both cases, measures at ages 7, 8, and 9 are used. Estimated are the following population components of the models: (a) $\alpha_k^C E(\theta^C(1) - \theta^C(0))$; (b) $\alpha_k^E E(\theta^E(1) - \theta^E(0))$; (c) $\alpha_k^I E(\theta^A(1) - \theta^A(0))$; (d) $\alpha_k^E E(\theta^E(1) - \theta^E(0)) + \alpha_k^A E(\theta^A(1) - \theta^A(0))$; (e) $\boldsymbol{\alpha}_k E(\boldsymbol{\theta}(1) - \boldsymbol{\theta}(0))$; (f) τ_k, where "C" stands for "Cognition", "E" stands for "Externalizing Behavior", "A" stands for "Academic Motivation". One-sided bootstrap p-values are reported. p-values below 0.1 are in bold italics. The number of bootstrap iterations is 1000. Stars denote significance levels: ** - 5% and * - 10%. $\boldsymbol{X}$ includes three background variables at entry: mother's employment, father's presence in the household, and family's SES. "CAT total" denotes the California Achievement Test total score.

Figure L.3: Decompositions of Treatment Effects, Cognition Measured by IQs versus Achievement Scores



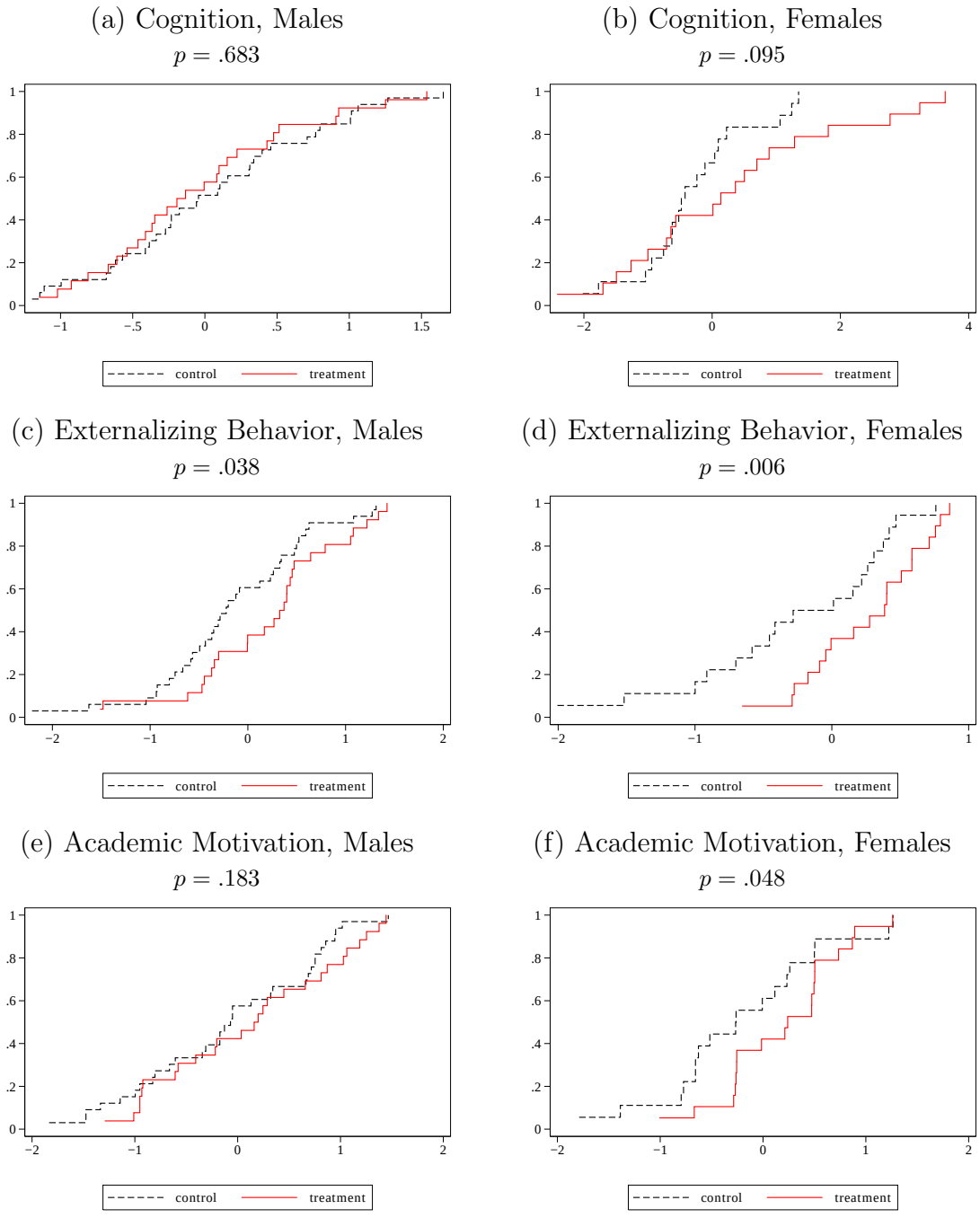
Notes: We calculate these comparisons for a reduced sample size, for which both IQ and CAT measures are non-missing, which alters the full sample IQ estimates. The total treatment effect is normalized to 100%. One-sided p-values are shown above each component in each outcome. (1) "IQ" denotes a model where cognition is measured by IQ at ages 7, 8, and 9. (2) "CAT" denotes a model where cognition is measured by the California Achievement Test. In both cases, measures at ages 7, 8, and 9 are used. The figure is a slightly simplified visualization of Tables L.12 and L.13 of Web Appendix L: small and statistically insignificant contributions of the opposite sign are set to zero.

Figure L.4: Decompositions of Treatment Effects, Factor Scores versus MLE



Notes: The total treatment effect is normalized to 100%. For each component of each outcome, one-sided p-values are shown above the corresponding component. "FACTOR" denotes the three-step estimation method using factor scores as described in the main paper. "MLE" denotes a one-step maximum likelihood estimation method where both measurement system and outcome equation are estimated simultaneously. The figure is a slightly simplified visualization of Tables L.5 and L.6. Small and statistically insignificant contributions of the opposite sign are set to zero.

Figure L.5: CDFs of Factor Scores



Notes: Cumulative distribution functions of Bartlett (1937) factor scores are shown. (See the discussion of the Bartlett procedure in Web Appendix F.) Numbers above the graphs are one-sided bootstrap p-values testing the equality of factor score means for the treatment and control groups. Graphs with corresponding kernel densities are shown in Figure 5 of the paper. Scores are defined leased on dedicated measures presented in Figure 2 of the paper.

Table L.14: Regression Coefficients used for Decompositions, Males

Outcome	Statistic	Cognition effect[a]	Externalizing Behavior[b]	Academic Motivation[c]	Mother Working[d]	Father Presence[e]	Socio-economic Status[f]	Available Observations
CAT total at age 14, end of grade 8 (+)	effect	.819 **	-.203	.700	-.156	.299	-.010	45
	p- value	***.000***	.845	***.000***	.597	.256	.964	
# of misdemeanor arrests, age 27 (-)	effect	-.259	-1.226 **	-.152	-1.413	.582	-.073	59
	p- value	.359	***.028***	.367	.174	.508	.799	
# of felony arrests, age 27 (-)	effect	-.618	-1.333 **	.219	.529	.855	-.291	59
	p- value	.235	***.023***	.557	.681	.298	.316	
# of adult arrests (misd.+fel.), age 27 (-)	effect	-.876	-2.559 **	.067	-.884	1.437	-.364	59
	p- value	.251	***.014***	.549	.639	.315	.486	
Monthly income, age 27 (+)	effect	.970 **	.698 **	-.257	.681	-.429	-.027	55
	p- value	***.038***	***.046***	.670	.256	.240	.887	
Use tobacco, age 27 (-)	effect	-.179	-.332 **	.159	.084	.168	.012	57
	p- value	.121	***.001***	.847	.606	.218	.773	
# of misdemeanor arrests, age 40 (-)	effect	-.620	-2.424 *	.196	-.990	-.089	-.866	59
	p- value	.383	***.087***	.501	.753	.967	.223	
# of felony arrests, age 40 (-)	effect	-.628	-1.755 **	.293	.793	1.623	-.701 *	59
	p- value	.266	***.014***	.570	.613	.112	***.085***	
# of adult arrests (misd.+fel.), age 40 (-)	effect	-1.248	-4.180 **	.489	-.197	1.534	-1.567	59
	p- value	.327	***.039***	.525	.974	.557	.129	
# of lifetime arrests, age 40 (-)	effect	-1.100	-4.740 **	.239	-.552	1.893	-1.629	59
	p- value	.359	***.030***	.519	.906	.526	.160	
Employed, age 40 (+)	effect	.277 **	.230 **	-.270	.368 **	-.212	.009	54
	p- value	***.012***	***.011***	.991	***.023***	.125	.827	

Notes: Regression coefficients for factor scores in Equation (5) are shown with one-sided p-values in parentheses. (+) and (-) denote the sign of the total treatment effect on the corresponding variable. Estimates are corrected based on the bias-correcting procedure described in Equation (A4). "CAT total" denotes the California Achievement Test total score. See Tables L.14 and L.15 of Web Appendix L for more detailed versions of this table. Stars denote significance levels: *** - 1%, ** - 5%, and * - 10%.

Table L.15: Regression Coefficients used for Decompositions, Females

Outcome	Statistic	Cognition effect[a]	Externalizing Behavior[b]	Academic Motivation[c]	Mother Working[d]	Father Presence[e]	Socio-economic Status[f]	Available Observations
CAT total, age 8 (+)	effect	.219 **	-.134	.689 **	.221	.125	-.004	35
	p-value	***.039***	.729	***.000***	.654	.787	.991	
CAT total, age 14 (+)	effect	.154	-.448	.899 **	.410	-.140	.046	31
	p-value	.113	.931	***.001***	.485	.792	.880	
Any special education, age 14 (-)	effect	-.041	.119	-.209 *	-.469 **	.132	-.004	37
	p-value	.273	.759	***.064***	***.018***	.410	.946	
Mentally impaired at least once, age 19 (+)	effect	-.039	.227	-.308 **	-.274	-.092	.007	33
	p-value	.283	.948	***.008***	.145	.567	.928	
# of misdemeanor violent crimes, age 27 (-)	effect	.083	-1.080 **	.150	-.629	.795	.161	37
	p-value	.778	***.043***	.700	.283	.258	.433	
# of felony arrests, age 27 (+)	effect	.021	-.451 *	.140	-.037	.161	-.030	37
	p-value	.609	***.053***	.808	.895	.594	.690	
Jobless for more than 1 year, age 27 (-)	effect	.139	.048	-.465 **	-.084	.043	-.044	36
	p-value	.920	.608	***.003***	.747	.893	.643	
Ever tried drugs other than alcohol or weed, age 27 (-)	effect	-.043	-.146	.122	-.026	-.069	.027	34
	p-value	.201	.144	.854	.838	.568	.529	
# of misdemeanor violent crimes, age 40 (-)	effect	.084	-1.078 **	.081	-.557	.667	.060	37
	p-value	.774	***.043***	.592	.339	.351	.798	
# of felony arrests, age 40 (-)	effect	.047	-.589 **	.078	-.150	.217	-.027	37
	p-value	.704	***.014***	.643	.608	.490	.772	
# of lifetime violent crimes, age 40 (-)	effect	.096	-1.220 **	.165	-.664	.789	.058	37
	p-value	.807	***.023***	.704	.255	.269	.819	
Months in all marriages, age 40 (+)	effect	21.748	13.591	10.453	47.857	-21.534	-19.348	36
	p-value	.111	.289	.280	.180	.490	.261	

Notes: Regression coefficients for factor scores in Equation (5) are shown with one-sided p-values in parentheses. (+) and (-) denote the sign of the total treatment effect on the corresponding variable. Estimates are corrected based on the bias-correcting procedure described in Equation (A4). "CAT total" denotes the California Achievement Test total score. See Tables L.14 and L.15 of Web Appendix L for more detailed versions of this table. Stars denote significance levels: *** - 1%, ** - 5%, and * - 10%.

Table L.16: Testing for Treatment Effects on Cognitive Measures (One-sided p-Values)

	Age										
	At entry	3	4	5	6	7	8	9	10	11	14
Males											
IQ Tests											
Stanford Binet	0.143	—	***0.000***	***0.001***	***0.020***	0.156	0.697	0.769	0.896	—	—
Leiter	—	0.174	***0.001***	***0.010***	0.510	0.749	0.785	0.151	—	—	—
PPVT	—	***0.069***	***0.015***	***0.001***	0.128	0.414	0.153	0.430	—	—	—
ITPA	—	0.214	—	***0.000***	0.250	0.535	0.304	0.470	—	—	—
Achivement Test											
CAT	—	—	—	—	—	0.371	0.270	0.195	0.158	0.359	***0.056***
Females											
IQ Tests											
Stanford Binet	0.391	—	***0.000***	***0.002***	***0.055***	***0.031***	***0.051***	***0.086***	***0.024***	—	—
Leiter	—	***0.001***	***0.000***	***0.001***	***0.005***	***0.016***	***0.044***	***0.006***	—	—	—
PPVT	—	***0.056***	***0.003***	***0.002***	0.133	0.104	0.429	0.294	—	—	—
ITPA	—	***0.060***	—	***0.000***	***0.061***	***0.058***	***0.057***	***0.025***	—	—	—
Achivement Test											
CAT	—	—	—	—	—	***0.027***	***0.031***	***0.041***	0.146	0.286	***0.003***

Notes: One-sided robust asymptotic p-values are shown for the treatment effects on cognitive measures conditional on $\boldsymbol{X}$ (working mother, present father, and family SES). For a detailed description of cognitive tests see Section B of the Web Appendix.

Table L.17: Testing for Treatment Effects on PBI Measures (One-sided p-Values)

item #	item description	males	females	item #	item description	males	females
1	Shows initiative	0.171	0.394	18	Requires continuous supervision	0.384	***0.002***
2	Blames others for troubles	***0.083***	***0.009***	19	Aggressive toward peers	0.125	***0.011***
3	Resistant to teacher	0.352	0.130	20	Disobedient	0.135	0.137
4	Alert and interested in school work	0.191	0.139	21	Steals	0.185	***0.030***
5	Attempts to manipulate adults	0.650	***0.019***	22	Friendly and well-received by other pupils	***0.054***	0.218
6	Appears depressed	0.460	***0.007***	23	Easily led into trouble	0.143	***0.011***
7	Learning retained well	0.372	***0.034***	24	Resentful of criticism or discipline	0.462	***0.043***
8	Absences or truancies	0.408	0.308	25	Hesitant to try, or gives up easily	0.363	***0.068***
9	Withdrawn and uncommunicative	0.277	0.540	26	Uninterested in subject matter	0.312	***0.035***
10	Completes assignments	0.488	***0.085***	27	Disrupts classroom procedures	0.197	***0.070***
11	Influences others toward trouble making	***0.071***	***0.059***	28	Swears or uses obscene words	***0.044***	0.114
12	Inappropriate personal appearance	0.377	0.385	29	Appears generally happy	0.549	***0.032***
13	Seeks constant reassurance	0.770	***0.081***	30	Poor personal hygiene	0.463	0.367
14	Motivated toward academic performance	0.612	***0.072***	31	Possessive of teacher	0.630	0.161
15	Impulsive	0.257	***0.007***	32	Teases or provokes students	***0.049***	0.175
16	Lying or cheating	***0.061***	***0.003***	33	Isolated, few or no friends	***0.062***	***0.097***
17	Positive concern for own education	0.294	0.191	34	Shows positive leadership	0.333	0.181

Notes: One-sided robust asymptotic p-values are shown for the treatment effects on PBI measures conditional on $\boldsymbol{X}$ (working mother, present father, and family SES). For a detailed description of PBI measures see Section C of the Web Appendix. Measures are averages over non-missing PBI items for ages 7, 8, and 9.

Table L.18: Testing Treatment Effects on YRS Measures (One-sided p-Values)

item #	item description	males	females
1	Social relationship with class mates	0.168	***0.029***
2	Social relationship with teacher	0.419	***0.083***
3	Level of verbal communication	0.151	0.622
4	Degree of imagination and creativity shown in handling materials and equipment	0.331	0.250
5	Level of academic readiness	0.121	0.102
6	Level of curiosity shown	0.498	0.505
7	Level of emotional adjustment	0.458	0.112
8	Prediction of future academic success	0.267	***0.048***
9	Degree of your desire to work with this child	0.116	0.123
10	Degree of trust of total environment	0.139	0.124
11	Direction of interest (introversion - extroversion)	0.526	0.483
12	Mother's degree of cooperation shown	0.733	0.914
13	Prediction of mother's future school relationship	0.721	0.929

Notes: One-sided robust asymptotic p-values are shown for the treatment effects on YRS measures conditional on $\boldsymbol{X}$ (working mother, present father, and family SES). For a detailed description of YRS measures see Section D of the Web Appendix. Measures are averages over non-missing YRS items for ages 7, 8, and 9.

Table L.19: Testing Treatment Effects on Various Indices (One-sided p-Values)

	males	females
Original Indicies		
PBI Indicies[(a)]		
Personal Behavior	***0.093***	***0.033***
Classroom Conduct	0.157	***0.013***
PBI Academic Motivation	0.325	***0.076***
PBI Socio-Emotional State	0.167	***0.073***
PBI Teacher Dependence	0.732	***0.097***
YRS Indicies[(b)]		
Academic Potenial	0.214	***0.086***
Verbal Skill	0.151	0.622
Social Development	0.332	0.114
Emotional Adjustment	0.229	0.105
Big Five Traits (PBI and YRS-based) - dedicated measures only[(c)]		
Openness	0.355	***0.064***
Conscientiousness	0.157	***0.017***
Extraversion	0.130	***0.064***
Agreeableness	0.125	***0.011***
Neuroticism	0.440	***0.014***
Big Five Traits (PBI and YRS-based) - all measures[(d)]		
Openness	0.295	***0.061***
Conscientiousness	0.156	***0.015***
Extraversion	***0.079***	***0.022***
Agreeableness	0.110	***0.016***
Neuroticism	0.278	***0.011***
Indicies constructed in this paper[(e)]		
Externalizing Behavior	***0.052***	***0.020***
Academic Motivation	0.221	0.147

Notes: One-sided robust asymptotic p-values are shown for the treatment effects on various indices measures conditional on $\boldsymbol{X}$ (working mother, present father, and family SES). All indices are unweighed averages of normalized PBI and YRS measures. The measures are averages over non-missing values of corresponding items at ages 7, 8, and 9. $^{(a)}$The PBI indices are defined in Table C.1. $^{(b)}$The YRS indices are defined in Table D.1. $^{(c)}$The indices are based on those PBI and YRS measures from Tables C.1 and D.1 representing only one trait (see traits in parentheses in Tables C.1 and D.1). $^{(d)}$These indices are similar to those described in (c), but now they are based on more measures, since we use not only measures representing one trait, but measures representing multiple traits. For instance, a measure denoted (A/C) is used twice: to represent Agreeableness (A), and to represent Conscientiousness (C). $^{(e)}$Personality indices constructed in this paper are defined in Table 2 of the main paper.

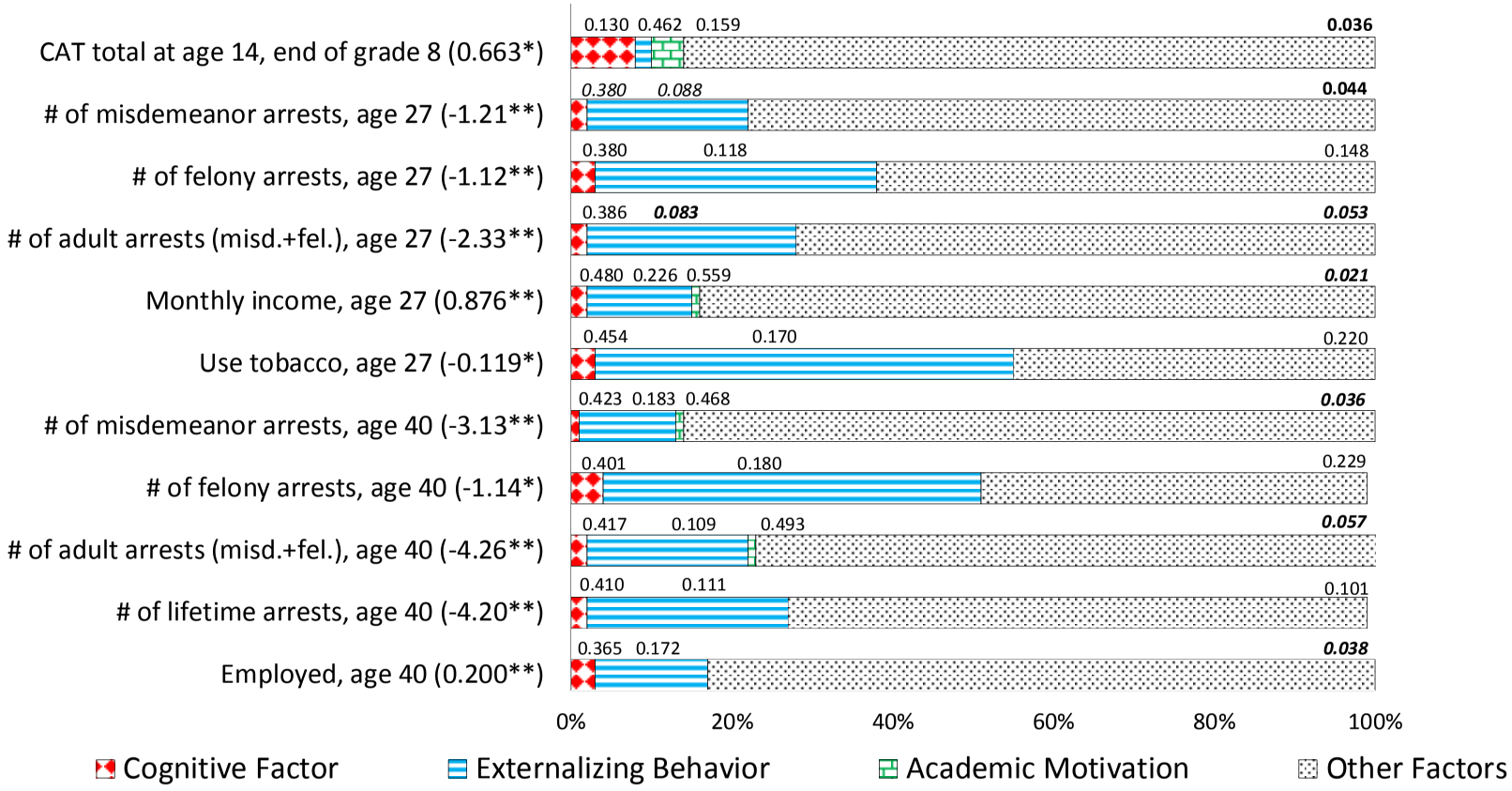


Figure L.6: Decompositions of Treatment Effects by Indices, Males

Note: The total treatment effects are shown in parentheses. Each bar represents the total treatment effect normalized to 100 percent. One-sided p-values are shown above each component of the decomposition. "CAT total" denotes California Achievement Test total score. Asterisks denote statistical significance: * – 10 percent level; ** – 5 percent level; *** – 1 percent level.

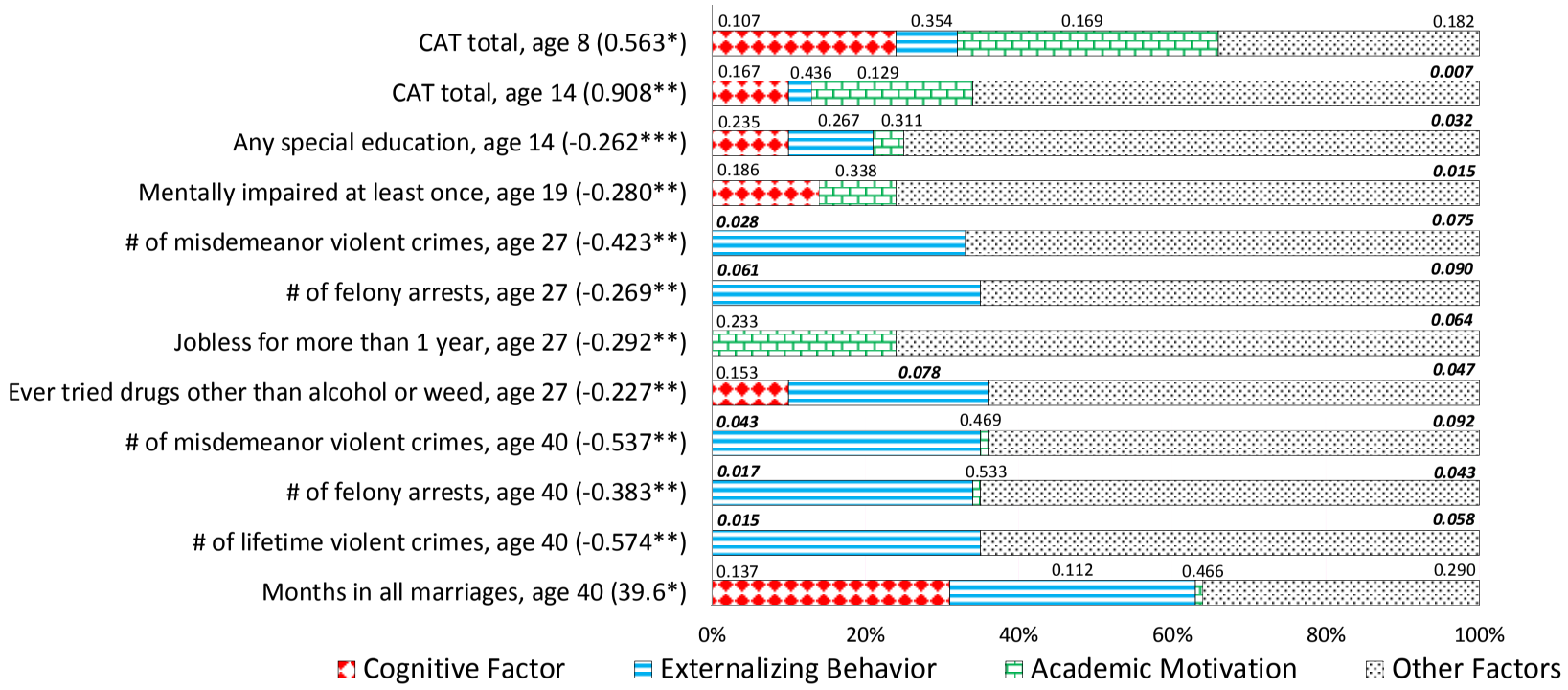


Figure L.7: Decompositions of Treatment Effects by Indices, Females

Note: The total treatment effects are shown in parentheses. Each bar represents the total treatment effect normalized to 100 percent. One-sided p-values are shown above each component of the decomposition. "CAT total" denotes California Achievement Test total score. Asterisks denote statistical significance: * – 10 percent level; ** – 5 percent level; *** – 1 percent level.

M Tests of the Validity of the Extracted Factor System

Our factor model imposes restrictions analogous to separability restrictions in demand analysis. In this appendix, we test the validity of the derived factor structure. In particular, we test if the measures excluded in the exploratory factor analysis have predictive power conditional on the extracted factors. Adopt a general notation for the outcome and measurement equations for this section to simplify exposition.

$$\text{(Outcome): } \boldsymbol{Y} = \boldsymbol{\tau}_0 + \boldsymbol{\tau_1} D + \boldsymbol{\alpha\theta} + \boldsymbol{\beta} X + \boldsymbol{\epsilon} \tag{M-1}$$

$$\text{(Measurement): } \boldsymbol{M} = \boldsymbol{\nu} + \boldsymbol{\varphi\theta} + \boldsymbol{\eta}. \tag{M-2}$$

Assume $\text{Dim}(\boldsymbol{M}) \gg \text{Dim}(\boldsymbol{\theta})$ so that it satisfies the Ledermann bound. (See Anderson and Rubin, 1956). We conduct two kinds of tests:

> Test I: **Conditional on extracted factors**, do unused components of $\boldsymbol{M}$ differ across $d = 0$ and $d = 1$ states?
>
> Test II: **Conditional on extracted factors**, do unused components of $\boldsymbol{M}$ predict $\boldsymbol{Y}$?

Evidence from both types of tests support the low-dimensional specification of equations derived from applying EFA that is used in the text.

Thus we run two types of regressions to test whether the unused measures dropped by EFA exhibit a treatment effect conditional on the extracted factors. First, we regress each of the unused measures on the treatment status indicator D, the estimated factors $\hat{\boldsymbol{\theta}}$, and background variables $\boldsymbol{X}$. Second, we create PBI and YRS indices (five in total) of the unused measures as defined by the Perry psychologists.[30] We run regressions analogous to those described for the unused measures using indices instead of each of the unused measures as dependent variables. In both types of regressions, we perform a joint hypothesis test to

[30] See Tables C.1 and D.1 of Web Appendices C and D for the description of the PBI and YRS scales. The indices used in this test procedure are unweighted averages of unused components of each scale.

see if all treatment coefficients are zero using the stepdown procedure of Romano and Wolf (2005) to avoid spurious p-values arising from testing multiple hypothesis. We also perform a χ^2-squared test for the hypothesis that conditioning on the extracted factors, all the unused measures show no treatment effect.[31] We apply a similar procedure to the indices of the unused measures. Adjusting for estimation error, we can reject neither of these joint null hypotheses (see Tables M.1 and M.2).

We also check whether conditional on the extracted factors, the unused measures explain outcomes. We run two types of regressions. First, we regress outcomes on each of the unused measures, the estimated factors, $\hat{\boldsymbol{\theta}}$, and background variables $\boldsymbol{X}$. For each outcome, we run 33 regressions, measure by measure, for all unused measures. We report p-values for the χ^2 test based on 1000 bootstrap draws. Second, instead of using individual measures, we use the same PBI and YRS indices of the unused measures as employed in the previous tests. As before, for both sets of tests, we correct p-values for the effect of testing multiple hypotheses using the stepdown procedure of Romano and Wolf (2005). We also perform a χ^2 test for the joint hypothesis that, conditioning on the extracted factors, all unused measures do not predict outcomes (see Table M.3). We find that neither unused measures nor indices based on the unused measures affect outcomes when extracted factors are controlled for. Results from these specification tests lend credibility to the factors extracted from the EFA approach.[32]

Tables M.1–M.2 test whether measures unused for model estimation show any treatment effect conditional on skills $\boldsymbol{\theta}$ and background variables $\boldsymbol{X}$ and show no such effects after adjusting for multiple hypothesis testing. In a similar fashion, Table M.3 shows that the unused measures do not affect life outcomes.

[31] The χ^2 test only weakly controls for the family wise error rate (FWER), while the stepdown procedure strongly controls the FWER. By "weak control" we mean that the probability of falsely reject a true hypothesis is below an adopted significance level only if all null hypotheses are true. By "strong control" we mean that the probability of falsely reject a true hypothesis is below an adopted significance level regardless of the number of true null hypotheses. See Lehmann and Romano, 2005.

[32] All test statistics are adjusted for the estimation error in creating $\hat{\boldsymbol{\theta}}$.

Table M.1: Testing Whether the Treatment Effect on the Unused Measures is Zero

Measures	Males				Female			
	effect	std. error	p-value	adjusted[(a)]	effect	std. error	p-value	adjusted[(a)]
PBI requires continuous supervision	-0.221	0.159	0.914	–[(b)]	0.628	0.249	***0.009***	0.401
PBI appears depressed	-0.010	0.235	0.518	–	0.731	0.326	***0.016***	–[(c)]
YRS prediction of future academic success	0.028	0.127	0.412	–	0.432	0.222	***0.031***	–
YRS social relationship with class mates	0.086	0.223	0.350	–	0.319	0.191	***0.054***	–
YRS level of academic readiness	0.171	0.148	0.126	–	0.267	0.173	***0.068***	–
PBI learning retained well	-0.060	0.168	0.639	–	0.190	0.127	***0.073***	–
PBI blames others for troubles	0.064	0.195	0.373	–	0.381	0.286	***0.096***	–
PBI impulsive	-0.121	0.224	0.704	–	0.308	0.237	0.101	–
PBI appears generally happy	0.017	0.245	0.473	–	0.322	0.262	0.114	–
PBI uninterested in subject matter	-0.103	0.165	0.734	–	0.254	0.218	0.127	–
PBI isolated, few or no friends	0.384	0.267	0.078	–	0.413	0.363	0.132	–
PBI easily led into trouble	0.005	0.174	0.489	–	0.263	0.232	0.133	–
YRS social relationship with teacher	-0.065	0.279	0.591	–	0.240	0.234	0.157	–
PBI seeks constant reassurance	-0.398	0.234	0.953	–	0.306	0.363	0.203	–
PBI motivated toward academic performance	-0.314	0.120	0.994	–	0.084	0.130	0.262	–
YRS level of emotional adjustment	-0.164	0.202	0.790	–	0.162	0.252	0.263	–
PBI attempts to manipulate adults	-0.422	0.183	0.987	–	0.078	0.127	0.272	–
PBI shows positive leadership	0.082	0.225	0.359	–	0.144	0.242	0.278	–
YRS degree of trust of total environment	0.042	0.160	0.398	–	0.116	0.237	0.314	–
PBI completes assignments	-0.249	0.134	0.966	–	0.044	0.111	0.347	–
yrs degree of imagination and creativity shown	0.052	0.177	0.385	–	0.112	0.326	0.366	–
PBI resentful of criticism or discipline	-0.256	0.257	0.838	–	0.070	0.224	0.379	–
PBI possessive of teacher	-0.236	0.223	0.852	–	0.015	0.314	0.481	–
YRS level of curiosity shown	-0.034	0.227	0.558	–	-0.012	0.187	0.526	–
PBI withdrawn and uncommunicative	0.195	0.229	0.200	–	-0.044	0.341	0.550	–
PBI positive concern for own education	-0.122	0.124	0.836	–	-0.034	0.197	0.567	–
PBI inappropriate personal appearance	0.074	0.212	0.364	–	-0.060	0.327	0.572	–
PBI friendly and well-received by other pupils	0.238	0.211	0.132	–	-0.048	0.183	0.604	–
PBI poor personal hygiene	0.000	0.219	0.501	–	-0.083	0.297	0.609	–
PBI absences or truancies	-0.040	0.214	0.574	–	-0.060	0.192	0.622	–
YRS level of verbal communication	0.319	0.238	0.093	–	-0.097	0.236	0.658	–
PBI resistant to teacher	-0.211	0.185	0.871	–	-0.140	0.145	0.829	–
PBI disobedient	-0.086	0.124	0.754	–	-0.193	0.075	0.992	–
Joint Test[(d)]			0.352				0.221	

Notes: Unused measures are those measures of skills are in the data but are not used to estimate our model. We test whether treatment affects the unused measures of skills at ages 7–9 conditional on factors $\boldsymbol{\theta}$ and background variables $\boldsymbol{X}$. We then correct p-values for multiple hypothesis testing based on the Romano and Wolf (2005) stepdown method. $^{(a)}$Stepdown p-values are presented using a procedure outlined in the appendix of Heckman et al. (2010*b*), which is based on the Romano and Wolf (2005) method. This procedure corrects for multiple hypothesis testing. Failure to do so for a large set of hypotheses, such as the set analyzed in the paper, can produce spuriously significant findings. See the discussion in Romano, Shaikh and Wolf (2010). We use bootstrap based on 1000 draws for the stepdown procedure. Since the stepdown adjustment is necessary for females only, we order hypotheses so that p-values for females are ordered from the smallest to the largest (not the case for males). $^{(b)}$As all single-hypotheses tests cannot be rejected, there is no need to carry out the stepdown procedure for these sets of measures. $^{(c)}$As the adjusted p-value for the previous step showed no statistically significant result, we stop the stepdown procedure and conclude that we cannot reject all subsequent tests. $^{(d)}$We report the p-value for the joint test that all regression coefficients associated with treatment status are zero. We use the χ^2 test based on 1000 bootstrap draws.

Table M.2: Testing Whether the Treatment Effect on Indices Based on the Unused Measures is Zero[(a)]

Skill	Effect	std. error	*p* -value	adjusted[(b)]
Males				
PBI Socioemotional State	0.209	0.225	0.179	–[(c)]
YRS Academic Potential	0.089	0.124	0.239	–
YRS Social Development	-0.004	0.221	0.508	–
YRS Emotional Adjustment	-0.067	0.171	0.652	–
PBI Teacher Dependence	-0.356	0.219	0.945	–
Joint Test[(d)]			0.443	
Females				
YRS Academic Potential	0.286	0.190	***0.072***	0.180
PBI Socioemotional State	0.347	0.298	0.126	–[(e)]
YRS Social Development	0.204	0.191	0.147	–
YRS Emotional Adjustment	0.152	0.231	0.258	–
PBI Teacher Dependence	0.181	0.352	0.306	–
Joint Test[(d)]			0.287	

Notes: [(a)]Unused measures are those measures of skills are in the data but are not used to estimate our model. The indices used in this test procedure are unweighed averages of components of each scale as used by the Perry psychologists for those scales with components unused in estimating the model. We test whether treatment affects the indices conditional on factors $\boldsymbol{\theta}$ and background variables $\boldsymbol{X}$. We then correct p-values for multiple hypothesis testing based on the Romano and Wolf (2005) stepdown method. [(b)]Stepdown p-values are presented using a procedure outlined in the appendix of Heckman et al. (2010*b*), which is based on the Romano and Wolf (2005) method. We use bootstrap based on 1000 draws for the stepdown procedure. [(c)]As all single-hypotheses tests cannot be rejected, there is no need to carry out the stepdown procedure for these sets of measures. [(d)]We report the p-value for the joint test that all regression coefficients associated with treatment status are zero. We use the χ^2 test based on 1000 bootstrap draws. [(e)]As the adjusted p-value for the previous step showed no statistically significant result, we stop the stepdown procedure and conclude that we cannot reject all subsequent tests.

Table M.3: Testing Whether the Unused Measures Have No Effect on Outcomes

Outcomes	Unused Measures[a]		Indices[b]	
	p-value[c]	adjusted[d]	p-value[c]	adjusted[d]
Males				
# of felony arrests, age 40 (-)	.103	–[e]	.242	–[e]
# of adult arrests (misd.+fel.), age 40 (-)	.121	–	.586	–
# of lifetime arrests, age 40 (-)	.125	–	.513	–
# of felony arrests, age 27 (-)	.133	–	.135	–
Monthly income, age 27 (+)	.173	–	.550	–
# of misdemeanor arrests, age 40 (-)	.196	–	.750	–
# of adult arrests (misd.+fel.), age 27 (-)	.235	–	.495	–
CAT total at age 14, end of grade 8 (+)	.345	–	.696	–
Use tobacco, age 27 (-)	.359	–	.757	–
# of misdemeanor arrests, age 27 (-)	.468	–	.863	–
Employed, age 40 (+)	.531	–	.683	–
Joint χ^2 Test[g]	.180		.658	
Females				
Jobless for more than 1 year, age 27 (-)	***.074***	.384	.560	–[f]
CAT total, age 8 (+)	***.086***	.377	**.028**	0.161
Any special education, age 14 (-)	*.174*	–[f]	.140	–
# of misdemeanor violent crimes, age 40 (-)	*.301*	–	.215	–
# of misdemeanor violent crimes, age 27 (-)	*.322*	–	.213	–
Mentally impaired at least once, age 19 (+)	*.356*	–	.659	–
Months in all marriages, age 40 (+)	*.392*	–	.235	–
Ever tried drugs other than alcohol or weed, age	*.419*	–	.460	–
# of lifetime violent crimes, age 40 (-)	*.424*	–	.342	–
# of felony arrests, age 27 (+)	*.449*	–	.373	–
CAT total, age 14 (+)	*.456*	–	.322	–
# of felony arrests, age 40 (-)	*.600*	–	.539	–
Joint χ^2 Test[g]	*.330*		.255	

Notes: [a] Unused measures are those measures of skills that exist in the data but are not used for model estimation. We test whether the unused measures of skills at ages 7–9 predict a life outcome conditional on factor and control variables that are used in the model.

[b] Similar to the approach used in footnote (a), we test whether indices based on the unused measures predict outcomes conditional on $\boldsymbol{\theta}$ and $\boldsymbol{X}$. The indices used in this procedure are unweighed averages of components of each scale as they are defined by the Perry psychologists among the scales with components unused in estimating the model.

[c] To obtain a p-value for a particular outcome, we regress that outcome on each of the 33 unused measures conditional on $\boldsymbol{\theta}$ and $\boldsymbol{X}$. We then perform the χ^2 test based on 1000 bootstrap draws to test if all unused measures have no effect (similar for indices). Outcomes are listed in the order of corresponding p-values for "unused measures" panel.

[d] Stepdown p-values are presented using the procedure outlined in the appendix of Heckman et al. (2010*b*), which is based on the Romano and Wolf (2005) method.

[e] As the unadjusted p-values show no statistical significance, there is no need to carry out the stepdown procedure for these sets of measures.

[f] As the adjusted p-value for the previous steps of the stepdown procedure showed no statistically significant result, we can conclude that we cannot reject all subsequent hypotheses.

[g] We report the p-value for the joint test. We use the χ^2 test based on 1000 bootstrap draws.

References

Abbring, Jaap H., and James J. Heckman. 2007. "Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation." In *Handbook of Econometrics*. Vol. 6B, , ed. J. Heckman and E. Leamer, 5145–5303. Amsterdam:Elsevier.

Anderson, T.W., and Herman Rubin. 1956. "Statistical Inference in Factor Analysis." In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 5, , ed. Jerzey Neyman, 111–150. Berkeley:University of California Press.

Arthur, Grace. 1952. *The Arthur Adaptation of The Leiter International Performance Scale.* Washington D.C.:The Psychological Service Center Press.

Bartlett, M. S. 1937. "The Statistical Conception of Mental Factors." *British Journal of Psychology*, 28(1): 97–104.

Becker, Kirk A. 2003. "History of the Stanford-Binet Intelligence Scales: Content and Psychometrics." Riverside Publishing Stanford-Binet Intelligence Scales, Fifth Edition Assessment Service Bulletin 1, Itasca, IL.

Bentler, P.M. 1990*a*. "Comparative Fit Indexes in Structural Models." *Psychological Bulletin*, 107(2): 238 – 246.

Bentler, P. M. 1990*b*. "Fit Indexes, Lagrange Multipliers, Constraint Changes and Incomplete Data in Structural Models." *Multivariate Behavioral Research*, 25(2): 163–172.

Bodrova, Elena, and Deborah J. Leong. 2001. *Tools of the Mind: A case study of implementing the Vygotskian approach in American early childhood and primary classrooms.* Geneva:International Bureau of Education, UNESCO.

Bolck, Annabel, Marcel Croon, and Jacques Hagenaars. 2004. "Estimating Latent Structure Models with Categorical Variables: One-Step Versus Three-Step Estimators." *Political Analysis*, 12(1): 3–27.

Borghans, Lex, Angela L. Duckworth, James J. Heckman, and Bas ter Weel. 2008. "The Economics and Psychology of Personality Traits." *Journal of Human Resources*, 43(4): 972–1059.

Borghans, Lex, Bart H. H. Golsteyn, James J. Heckman, and John Eric Humphries. 2011. "IQ, Achievement, and Personality." Unpublished manuscript, University of Maastricht and University of Chicago (revised from the 2009 version).

Browne, Michael W. 2001. "An Overview of Analytic Rotation in Exploratory Factor Analysis." *Multivariate Behavioral Research*, 36: 111–150.

Browne, Michael W., and Robert Cudeck. 1992. "Alternative Ways of Assessing Model Fit." *Sociological Methods Research*, 21(2): 230–258.

Brown, L. F., and J. A. Rice. 1967. "The Peabody Picture Vocabulary Test: validity for EMR's." *American Journal of Mental Deficiency*, 71(6): 901–903.

Carroll, John. 1953. "An analytical solution for approximating simple structure in factor analysis." *Psychometrika*, 18: 23–38. 10.1007/BF02289025.

Cattell, Raymond B. 1966. "The Scree Test for the Number of Factors." *Multivariate Behavioral Research*, 1(2): 245–276.

Crawford, Charles, and George Ferguson. 1970. "A general rotation criterion and its use in orthogonal rotation." *Psychometrika*, 35: 321–332. 10.1007/BF02310792.

Croon, Marcel A. 2002. "Using Predicted Latent Scores in General Latent Structure Models." In *Latent Variable and Latent Structure Models.* , ed. G. A. Marcoulides and I. Moustaki, 195–223. NJ:Lawrence Erlbaum Associates, Inc.

Cunha, Flavio, and James J. Heckman. 2008. "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Journal of Human Resources*, 43(4): 738–782.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." Forthcoming, *Econometrica*.

Dewey, John. 1997. *Experience and Education.* New York:Free Press.

Dunn, Lloyd M. 1965. *Peabody Picture Vocabulary Test.* Minneapolis, MN:American Guidance Service.

Fabrigar, Leandre R., Duane T. Wegener, Robert C. MacCallum, and Erin J. Strahan. 1999. "Evaluating the Use of Exploratory Factor Analysis in Psychological Research." *Psychological Methods*, 4: 272–299.

Glass, Gene V., and Kenneth D. Hopkins. 1995. *Statistical Methods in Education and Psychology.* . 3 ed., Boston, MA:Allyn and Bacon.

Gorsuch, R. L. 2003. "Handbook of psychology: Vol 2. Research methods in psychology." , ed. J. A. Schinka and W. F. Velicer, Chapter Factor Analysis, 143–164. Hoboken, NJ: Wiley.

Guttman, Louis. 1954. "Some necessary conditions for common-factor analysis." *Psychometrika*, 19: 149–161.

Heckman, James J., and Rodrigo Pinto. 2012. "Econometric Mediation Analysis." Unpublished manuscript, University of Chicago.

Heckman, James J., and Tim Kautz. 2012. "Hard evidence on soft skills." *Labour Economics*, 19(4): 451–464.

Heckman, James J., Seong Hyeok Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam Q. Yavitz. 2010*a*. "The Rate of Return to the HighScope Perry Preschool Program." *Journal of Public Economics*, 94(1-2): 114–128.

Heckman, James J., Seong Hyeok Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam Q. Yavitz. 2010*b*. "A Reanalysis of the HighScope Perry Preschool Program." First draft, September, 2006. Forthcoming, *Quantitative Economics*.

Himelstein, Philip. 1966. "Research with the Stanford-Binet, Form L-M: The Five Years." *Psychological Bulletin*, 65(3): 156–164.

Hohmann, Mary, David P. Weikart, and Ann S. Epstein. 2008. *Educating Young Children.* Ypsilanti, MI:High/Scope Press.

Horn, John L. 1965. "A rationale and test for the number of factors in factor analysis." *Psychometrika*, 30(2): 179–185.

Horst, Paul. 1965. *Factor analysis of data matrices.* New York:Holt, Rinehart and Winston.

Jennrich, R. I., and P. F. Sampson. 1966. "Rotation for simple loadings." *Psychometrika*, 31(3): 313–323.

Jennrich, Robert I. 2006. "Rotation to Simple Loadings Using Component Loss Functions: The Oblique Case." *Psychometrika*, 71: 173–191. 10.1007/s11336-003-1136-B.

Jöreskog, Karl G., and Dag Sörbom. 1986. *LISREL VI :analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods.* . 4th ed., Mooresville, IN:Scientific Software.

Kaiser, Henry F. 1960. "The application of electronic computers to factor analysis." *Educational and Psychological Measurement*, 20: 141–151.

Kaiser, Henry F. 1961. "A note on Guttman's lower bound for the number of common factors." *British Journal of Statistical Psychology,*, 14: 1–2.

Kennedy, W. A., W. Nelson, R. Lindner, H. Moon, and J. Turner. 1960. "The Ceiling of the New Stanford-Binet." *Journal of Clinical Psychology*, 17: 284–286.

Lehmann, E. L., and Joseph P. Romano. 2005. *Testing Statistical Hypotheses.* . Third ed., New York:Springer Science and Business Media.

Lu, Irene R. R., and D. Roland Thomas. 2008. "Avoiding and Correcting Bias in Score-Based Latent Variable Regression with Discrete Manifest Items." *Structural Equation Modelling*, 15: 462–490.

Mulaik, Stanley A. 1972. *The Foundations of Factor Analysis.* McGraw-Hill (New York).

Onatski, Alexei. 2009. "Testing hypotheses about the number of factors in large factor models." *Econometrica*, 77(5): 14471479.

Piaget, Jean, and Bärbel Inhelder. 2000. *The Psychology of the Child.* New York:Basic Books.

Romano, Joseph P., and Michael Wolf. 2005. "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing." *Journal of the American Statistical Association*, 100(469): 94–108.

Romano, Joseph P., Azeem M. Shaikh, and Michael Wolf. 2010. "Hypothesis Testing in Econometrics." *Annual Review of Economics*, 2: 75–104.

Sattler, Jerome M. 1965. "Analysis of functions of the 1960 Stanford-Binet Intelligence Scale, form L-M." *Journal of Clinical Psychology*, 21(2): 115–232.

Schweinhart, Lawrence J., Helen V. Barnes, and David P. Weikart. 1993. *Significant Benefits: The High-Scope Perry Preschool Study Through Age 27.* Ypsilanti, MI:High/Scope Press.

Skrondal, Anders, and Petter Laake. 2001. "Regression among Factor Scores." *Psychometrika*, 66(4): 563–576.

Steiger, James H. 1990. "Structural Model Evaluation and Modification: An Interval Estimation Approach." *Multivariate Behavioral Research*, 25(2): 173.

Sylva, Kathy. 1997. "The Quest for Quality in Curriculum." In *Lasting Differences: The High/Scope Preschool Curriculum Comparison Study through Age 23.* , ed. L. J. Schweinhart and D. P. Weikart, 89–93. Ypsilanti:High/Scope Press.

Taylor, L. J. 1975. "The Peabody Picture Vocabulary Test: What does it measure?" *Perceptual and Motor Skills*, 41: 777–778.

Terman, Lewis Madison, and Maud A. Merrill. 1937. *Measuring Intelligence.* Boston:Houghton Mifflin.

Terman, Lewis Madison, and Maud A. Merrill. 1960. *Stanford-Binet Intelligence Scale: Manual for the Third Revision Form L-M.* Boston:Houghton Mifflin.

Thompson, Bruce. 2004. *Exploratory and confirmatory factor analysis.* Washington, DC: American Psychological Association.

Thurstone, L. L. 1935. *The Vectors of Mind: Multiple-Factor Analysis for the Isolation of Primary Traits.* Chicago, IL:University of Chicago Press.

Thurstone, L. L. 1947. *Multiple Factor Analysis.* Chicago, IL:University of Chicago Press.

Tiegs, E. W., and W. W. Clark. 1971. *California Achievement Tests.* Monterey Park, CA:McGraw-Hill for California Test Bureau.

Tucker, Ledyard, and Charles Lewis. 1973. "A reliability coefficient for maximum likelihood factor analysis." *Psychometrika*, 38: 1–10. 10.1007/BF02291170.

Vinter, Robert D., Rosemary C. Sarri, Darrel J. Vorwaller, and Walter E. Shafer. 1966. *Pupil Behavior Inventory: A Manual for Administration and Scoring.* Ann Arbor, MI:Campus Publishers.

Vygotsky, Lev S. 1986. *Thought and Language.* Cambridge, MA:MIT Press.

Wade, Teresa Hartung. 1978. "A Comparison of the Stanford-Binet Intelligence Scale and the McCarthy Scales of Children's Abilities with Preschool Children." *Psychology in the Schools*, 15(4): 468–472.

Wansbeek, Tom J., and Erik Meijer. 2000. *Measurement error and latent variables in econometrics. Advanced Textbooks in Economics, 37.* 1 ed., New York:Elsevier.

Weikart, David P., James T. Bond, and J. T. McNeil. 1978. *The Ypsilanti Perry Preschool Project: Preschool Years and Longitudinal Results Through Fourth Grade.* Ypsilanti, MI:Monographs of the High/Scope Educational Research Foundation.

Yates, A. 1987*a*. *Multivariate exploratory data analysis: A perspective on exploratory factor analysis.* Albany:State University of New York Press.

Yates, Allen. 1987*b*. *Multivariate exploratory data analysis: A perspective on exploratory factor analysis.* Albany, NY, US: State University of New York Press.

Zwick, William R., and Wayne F. Velicer. 1986. "Comparison of five rules for determining the number of components to retain." *Psychological Bulletin*, 99(3): 432–442.