American Economic Review 2013, 103(6): 2052–2086
http://dx.doi.org/10.1257/aer.103.6.2052

Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes[†]

By James Heckman, Rodrigo Pinto, and Peter Savelyev*

A growing literature establishes that high quality early childhood interventions targeted toward disadvantaged children have substantial impacts on later life outcomes. Little is known about the mechanisms producing these impacts. This paper uses longitudinal data on cognitive and personality skills from an experimental evaluation of the influential Perry Preschool program to analyze the channels through which the program boosted both male and female participant outcomes. Experimentally induced changes in personality skills explain a sizable portion of adult treatment effects. (*JEL* I21, I24, I28, J13, J24)

A growing literature establishes that early childhood environments substantially impact later life outcomes (e.g., Knudsen et al. 2006, Heckman 2008, and Almond and Currie 2011). Less is known about the channels through which early environments operate to produce their long-term effects. This paper examines the sources of the success of the Perry Preschool program, a flagship early childhood intervention in the United States.[1]

The Perry program was a randomized trial that targeted disadvantaged, low IQ African American children ages three to four. After two years, all participants left

* Heckman: Department of Economics, University of Chicago, 1126 E. 59th St., Chicago, IL 60637, University College Dublin, and American Bar Foundation (e-mail: jjh@uchicago.edu); Pinto: Department of Economics, University of Chicago, 1126 E. 59th St., Chicago, IL 60637 (e-mail: rodrig@uchicago.edu); Savelyev: Department of Economics, Vanderbilt University, PMB 351819, 2301 Vanderbilt Place, Nashville, TN 37235-1819, and Robert Wood Johnson Center for Health Policy, Meharry Medical College (e-mail: peter.savelyev@vanderbilt.edu). We thank three anonymous referees for helpful comments. Lena Malofeeva, then at the HighScope Foundation, worked with us in supplying the Perry data and interpreting it for us. We are deeply grateful for her assistance and the cooperation of the HighScope Foundation. Versions of this paper were presented at numerous conferences and seminars starting in 2006 that are listed in the online Appendix. 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, Maryclare Griffen, Jeff Grogger, Emir Kamenica, Costas Meghir, Jörn-Steffen Pischke, Devesh Raval, Brent Roberts, Cullen Roberts, Tino Sanandaji, Larry Schweinhart, Sandra Waxman, Ben Williams, and Junjian Yi for helpful comments. The paper was presented at a Public Policy and Economics Seminar at the Harris School, University of Chicago, October, 2009, attended by Diane Schatzenbach. 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 Pritzker Children's Initiative, the Buffett Early Childhood Fund, NICHD 5R37HD065072 and 5R01HD054702, the Human Capital and Economic Opportunity Global Working Group—an initiative of the Becker Friedman Institute for Research in Economics—funded by 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. The views expressed in this paper are those of the authors and not necessarily those of the funders or commentators named here.

[†] Go to http://dx.doi.org/10.1257/aer.103.6.2052 to visit the article page for additional materials and author disclosure statement(s).

[1] The formal name of the program is the "HighScope Perry Preschool Program" (see Schweinhart et al. 2005).

the program and entered the same public school. Data were collected for treatment and control groups through age 40.

Heckman et al. (2010a) and Conti et al. (2013) show that the Perry program significantly enhanced adult outcomes including education, employment, earnings, marriage, health, and participation in healthy behaviors, and reduced participation in crime.[2] We summarize many of these findings in Table 1. All treatment effects displayed there are statistically significant and survive adjustments for multiple hypothesis testing.[3] Heckman et al. (2010b) show that the internal rate of return to the program for both boys and girls is a statistically significant 6–10 percent per year—about the same as or larger than the historical return to equity.[4] Positive effects of the Perry program have become a cornerstone of the argument for preschool programs (e.g., Obama 2013). Currently, about 30 percent of all Head Start centers nationwide offer a version of the Perry curriculum (ICPSR 2010).[5]

Previous studies of Perry focus on estimating treatment effects and do not attempt to explain their sources.[6] This paper identifies the psychological skills changed by the Perry program and decomposes the treatment effects on adult outcomes displayed in Table 1 into components attributable to improvements in these skills.

The literature in the economics of education assumes the primacy of cognitive ability in producing successful lifetime outcomes (e.g., Hanushek and Woessmann 2008). From this perspective, the success of the Perry program is puzzling. Although the program initially boosted the IQs of participants, this effect soon faded. A few years after the program finished, there was no statistically significant difference in IQ between treatments and controls for males and only a borderline statistically significant difference for females (see Figure 1). Consistent with this evidence, we show negligible effects of increases in IQ in producing program treatment effects.

Although Perry did not produce long run gains in IQ, it did create persistent improvements in personality skills.[7] The Perry program substantially improved externalizing behaviors (aggressive, antisocial, and rule-breaking behaviors), which, in turn, improved a number of labor market outcomes and health behaviors and reduced criminal activities (see panels A and B of Figures 2 and 3).[8]

The program also enhanced academic motivation (see panels C and D of Figures 2 and 3), but the effect is primarily for girls.[9] This differential enhancement of endowments by gender helps to explain the positive treatment effects for education-related outcomes such as achievement tests and mental impairment for

[2] The small sample size of the Perry experiment (123 participants) has led some researchers to question the validity and relevance of its findings (e.g., Herrnstein and Murray 1994; Hanushek and Lindseth 2009). Heckman et al. (2010a) use a method of exact inference that is valid in small samples. They find that Perry treatment effects remain statistically significant even after accounting for multiple hypothesis testing and compromised randomization.

[3] One of the outcomes, the number of felony arrests for males at age 27, is borderline statistically significant at the 10 percent level.

[4] The historical post-World War II stock market rate of return to equity is 6.9 percent (DeLong and Magin 2009).

[5] Although not necessarily with the same quality of staff and background of participants as in the original program.

[6] See Weikart (1967); Weikart, Bond, and McNeil (1978); Berrueta-Clement et al. (1984); Schweinhart, Barnes, and Weikart (1993).

[7] See Heckman (2000) and Carneiro and Heckman (2003).

[8] Reduction in crime is a major benefit of the Perry program (Belfield et al. 2006; Heckman et al. 2010b).

[9] See Figure C.7 of the online Appendix for breakdowns by gender.

Table 1—Program Treatment Effects

	Treatment effect			Control group		Treatment group	
Variable	Effect	Effect size	p-value	Mean	Standard error	Mean	Standard error
Panel A. Males							
CAT total at age 14, end of grade 8	0.566*	0.652	(0.060)	0.000	(0.164)	0.566	(0.204)
Number of misdemeanor arrests, age 27	−1.21**	−0.363	(0.036)	3.03	(0.533)	1.82	(0.445)
Number of felony arrests, age 27	−1.12	−0.324	(0.101)	2.33	(0.554)	1.21	(0.342)
Number of adult arrests (misd.+fel.), age 27	−2.33**	−0.402	(0.024)	5.36	(0.927)	3.03	(0.734)
Monthly income, age 27	0.876**	0.607	(0.018)	1.43	(0.231)	2.31	(0.352)
Use tobacco, age 27	−0.119*	−0.236	(0.093)	0.538	(0.081)	0.419	(0.090)
Number of misdemeanor arrests, age 40	−3.13**	−0.372	(0.039)	8.46	(1.348)	5.33	(1.042)
Number of felony arrests, age 40	−1.14*	−0.266	(0.092)	3.26	(0.684)	2.12	(0.598)
Number of adult arrests (misd.+fel.), age 40	−4.26**	−0.373	(0.041)	11.7	(1.831)	7.46	(1.515)
Number of lifetime arrests, age 40	−4.20*	−0.346	(0.053)	12.4	(1.945)	8.21	(1.778)
Employed, age 40	0.200**	0.394	(0.024)	0.500	(0.085)	0.700	(0.085)
Sample size	72			39		33	
Panel B. Females							
CAT total, age 8	0.565*	0.614	(0.062)	0.000	(0.196)	0.565	(0.223)
CAT total, age 14	0.806**	0.909	(0.014)	0.000	(0.209)	0.806	(0.204)
Any special education, age 14	−0.262**	−0.514	(0.025)	0.462	(0.100)	0.200	(0.082)
Mentally impaired at least once, age 19	−0.280**	−0.569	(0.017)	0.364	(0.105)	0.083	(0.058)
Number of misdemeanor violent crimes, age 27	−0.423**	−0.292	(0.032)	0.423	(0.284)	0.000	(0.000)
Number of felony arrests, age 27	−0.269**	−0.325	(0.021)	0.269	(0.162)	0.000	(0.000)
Jobless for more than 1 year, age 27	−0.292*	−0.573	(0.071)	0.542	(0.104)	0.250	(0.090)
Ever tried drugs other than alcohol or weed, age 27	−0.227**	−0.530	(0.045)	0.227	(0.091)	0.000	(0.000)
Number of misdemeanor violent crimes, age 40	−0.537**	−0.364	(0.016)	0.577	(0.289)	0.040	(0.040)
Number of felony arrests, age 40	−0.383**	−0.425	(0.028)	0.423	(0.177)	0.040	(0.040)
Number of lifetime violent crimes, age 40	−0.574**	−0.384	(0.019)	0.654	(0.293)	0.080	(0.055)
Months in all marriages, age 40	39.6*	0.539	(0.076)	47.8	(15.015)	87.5	(18.853)
Sample size	51			26		25	

Notes: Statistics are shown for the outcomes analyzed in this paper. There are differences in treatment effects by gender although strong effects are found for both. "CAT total" denotes the California Achievement Test total score normalized to control mean zero and variance of one. Test statistics are corrected for the effect of multiple hypothesis testing and threats to validity (see Heckman et al. 2010a; Conti et al. 2013). The reported effect is the difference in means between treatment and control groups. The effect size is the ratio of the effect to the standard deviation of the control group. Stars denote statistical significance. Monthly income is adjusted to thousands of year-2006 dollars using annual national CPI.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

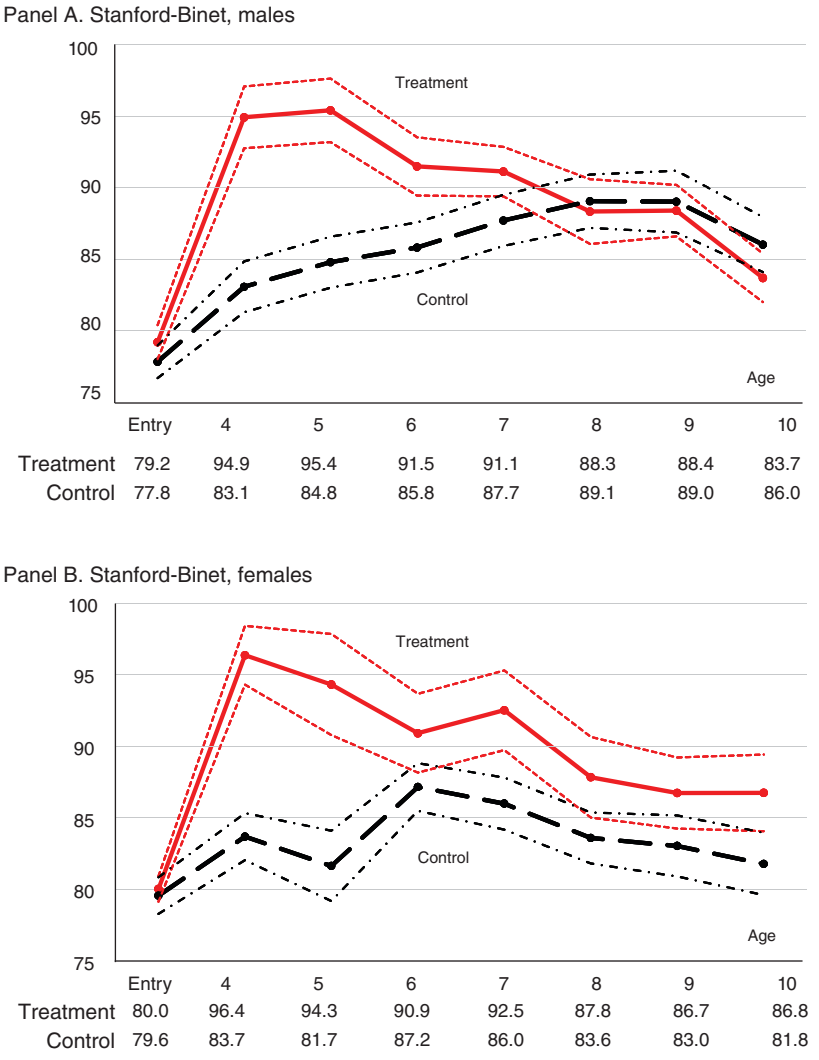

FIGURE 1. STANFORD-BINET IQ TEST SCORES BY GENDER AND TREATMENT STATUS

Notes: Bold lines display mean IQs. Fine lines represent standard errors for the corresponding means (one standard error above and below). For a detailed description of the cognitive measures and results for other IQ tests, see online Appendix B. Numbers below each chart are treatment and control mean test scores. See panels A–D of Figure B.6 of online Appendix B for comparable graphs on the Leiter and PPVT measures of IQ.

girls. Academic motivation is not significantly enhanced for boys, and plays no role in explaining their treatment effects.

While the Perry program did not boost long-term IQ, it did boost long-term achievement test scores (see panels E and F of Figures 2 and 3). The effect is stronger for girls, but also occurs for boys.[10] Achievement tests measure acquired

[10] See Figure B.5 in the online Appendix.

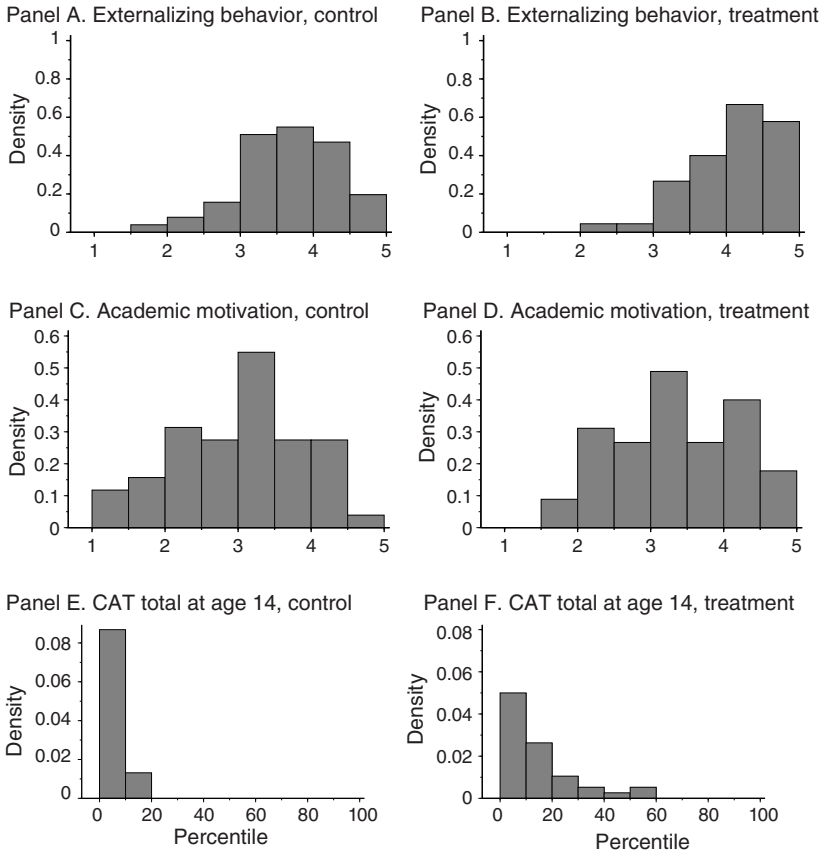

Figure 2. Histograms of Indices of Personality Skills and CAT Scores

Notes: Indices for externalizing behavior and academic motivation are unweighted averages of measures listed in Table 2. "CAT" is the California Achievement Test score expressed in percentiles of the general population distribution of the scores. See online Appendix B.4 for description of the CAT. The one-sided *p*-values for difference in means between treatments and controls are 0.001, 0.043, and 0.000 for externalizing behavior, academic motivation, and CAT scores respectively. Histograms are based on the pooled sample of males and females. See Figures C.6 and C.7 of online Appendix C and Figure B.5 of online Appendix B for the corresponding gender-specific figures.

knowledge, which is enhanced for children with better cognitive and personality skills. Enhanced personality skills promote learning, which, in turn, boosts achievement test scores.[11] This finding is consistent with recent evidence that 30–40 percent of the explained variance in achievement test scores across students is due to personality skills and not IQ.[12]

This paper contributes to an emerging literature on the economics of personality. Our demonstration of the powerful role of personality skills is in agreement with

[11] See Cunha and Heckman (2008) and Cunha, Heckman, and Schennach (2010) for evidence that personality skills boost acquisition of cognition as measured by achievement tests.

[12] Borghans et al. (2011a) show that achievement test scores are explained, in part, by both personality skills and IQ. See also Heckman and Kautz (2012).

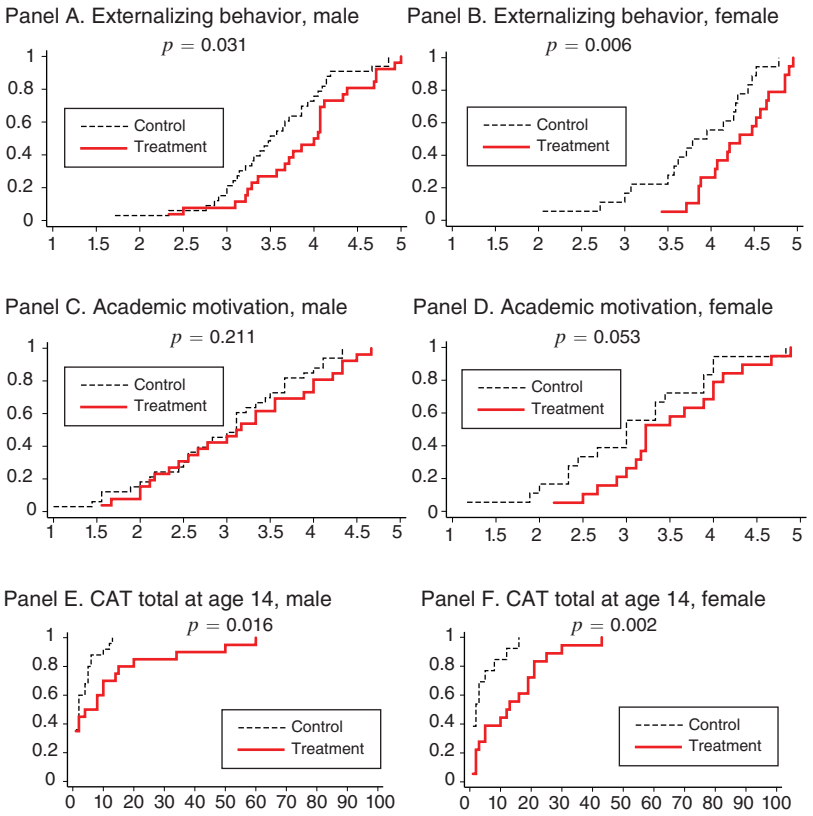

FIGURE 3. CUMULATIVE DISTRIBUTION FUNCTIONS OF INDICES OF PERSONALITY SKILLS AND CAT SCORES BY GENDER

Notes: Indices for externalizing behavior and academic motivation are unweighted averages of measures listed in Table 2. "CAT" is the California Achievement Test score expressed in percentiles of the general population distribution of the scores. Numbers above the charts are one-sided *p*-values testing the equality of means of the indices for the treatment and control groups.

a large body of evidence summarized in Borghans et al. (2008) and Almlund et al. (2011).[13]

Our analysis shows the benefits and limitations of social experiments. The Perry study generated experimentally determined treatment effects for outcomes and skills. However, knowledge that the program enhanced skills and improved a number of outcomes is not enough to establish that the improvement in measured skills *caused* the improvement in outcomes. Without further assumptions, data from the experiment do not determine the production function relating changes in skills to changes in adult outcomes. The program may also improve unmeasured skills. Changes in measured skills may simply proxy changes in unmeasured skills that affect outcomes. To address this issue, we supplement the treatment effects obtained from the experiment with an econometric model that estimates

[13] See also Bowles and Gintis (1976, 2001); Heckman, Stixrud, and Urzúa (2006); Segal (2008, forthcoming).

the relationship between outcomes and experimentally induced changes in measured skills. Our method accounts for measurement error and treatment-induced changes in unmeasured skills. Access to experimental data allows us to test some of its identifying assumptions. Evidence from a series of specification tests supports our econometric procedure.

The paper proceeds in the following way. Section I describes the Perry program and the experiment that evaluated it. Section II presents our econometric model. Section III discusses the variety of measures of psychological skills at our disposal and the need to create low-dimensional summaries of them. It explains how we construct summary measures and test for the validity of the constructed summaries. Section IV presents our analysis of the sources of the Perry treatment effects. Section V concludes. An online Appendix presents supplementary material.

I. The Perry Program: Design and Background

The Perry program targeted African American children with low IQs and socioeconomic status (Schweinhart and Weikart 1981). The experiment was conducted during the mid-1960s in the district of the Perry elementary school in Ypsilanti, Michigan. Children began the program at age three and were enrolled for two years.[14] Parents were disadvantaged as measured by their income and education. Roughly 47 percent of the children in the study did not have fathers present in the household at age three.

The 123 participants were randomized into treatment and control groups.[15] The Perry sample consists of 51 females (25 treatment and 26 control) and 72 males (33 treatment and 39 control). There was relatively little attrition: only 11 participants left the study by the time of the interview at age 40.[16]

The Perry curriculum is based on the principle of *active participatory learning*, in which children and adults are treated as equal partners in the learning process, and children engage with objects, people, events, and ideas.[17] Abilities to plan, execute, and evaluate tasks were fostered, as were social skills, including cooperation with others and resolution of interpersonal conflicts. The Perry curriculum has been interpreted as implementing the theories of Vygotsky (1986) in teaching self-control and sociability.[18] A widely implemented program based on these principles—Tools of the Mind—is designed to promote self-control.[19]

Sessions lasted 2.5 hours and were held five days a week during the school year. Teachers in the program, all of whom had bachelor's degrees (or higher) in

[14] The first entry cohort was enrolled for only one year of the program, beginning at age four.

[15] Heckman et al. (2010a) describe the protocol and develop statistical procedures for testing treatment effects which take into account the peculiarities of the Perry randomization protocol.

[16] Five control and two treatment group participants died; two control and two treatment group participants were missing.

[17] See online Appendix A for more information on the Perry curriculum.

[18] The curriculum of the Perry program was also grounded, in part, in the research on cognitive development by Jean Piaget (Piaget and Inhelder 2000) and in the progressive educational philosophy of John Dewey (Dewey 1997).

[19] See Tough (2009) for a popular exposition of the Tools of the Mind program. See Bodrova and Leong (2007) for a complete description of the Tools of the Mind program. Diamond et al. (2007) present a recent evaluation of the program that demonstrates that it enhanced self-control by participants. For a discussion of the Vygotskian foundations of the Perry program see Sylva (1997).

education, made weekly 1.5-hour home visits to treatment group mothers with the aim of involving them in the socio-emotional development of their children. The control group had no contact with the Perry program other than through annual testing and assessment (Weikart, Bond, and McNeil 1978).

Perry predates Head Start and had no competitors, so there was no control group contamination (see Schweinhart and Weikart 1981). All eligible parents enrolled their children in the program, so there was no issue of bias arising from noncompliance (Weikart, Bond, and McNeil 1978).

Numerous measures were collected annually from ages 3–15 on a variety of socioeconomic outcomes for treatment and control participants. There were three additional follow-ups at ages 19, 27, and 40. The Perry sample was representative of a particularly disadvantaged cohort of the African American population. About 16 percent of all African American children in the United States had family and personal attributes similar to those of Perry participants at the time when the Perry program was conducted.[20] The statistically significant treatment effects of the experiment for boys and girls survive rigorous adjustments for multiple hypothesis testing and compromises in the randomization protocol.[21]

II. Methodology

This paper explains the sources of the Perry treatment effects in terms of improvements in early measures of psychological skills broadly classified into cognitive and personality skills.[22] We first estimate treatment effects for these skills. We then estimate the relationship between skills and later life outcomes and decompose treatment effects for adult outcomes into components due to treatment-induced changes in different skills.[23]

To perform valid decompositions, we need to address two features of the Perry data. First, as previously noted, the randomized design of the Perry study allows us to identify the causal effect of the treatment on measured skills and on adult outcomes, but it does not directly allow us to identify the causal effect of increases

[20] Heckman et al. (2010a).

[21] Anderson (2008) adjusts test statistics for the Perry program treatment effects for the effects of multiple hypothesis testing. He claims that the program only affected girls. Heckman et al. (2010a) critically evaluate this conclusion and his procedures. They establish statistically significant program treatment effects for both boys and girls. Heckman et al. (2010b) show that the rate of return to the program is statistically significantly different from zero for both boys and girls.

[22] Throughout the paper we assume that the Perry program has either positive or no effect on outcomes and use one-sided *p*-values to test hypotheses. The literature shows that high-quality intervention programs targeting disadvantaged children generally show either beneficial or no effects from the program. For example, Gray and Klaus (1970); Lazar et al. (1982); Campbell and Ramey (1994, 1995); Yoshikawa (1995); and Reynolds et al. (2001) document beneficial effects of intervention programs targeting disadvantaged children. Barnett (1995) reviews a variety of early intervention programs and shows that there were mainly beneficial effects on children's development outcomes, although some programs had no treatment effects. He explains the lack of treatment effects as a consequence of the difference in program quality. Baker, Gruber, and Milligan (2008) investigate the effects of Quebec's universal childcare program and find a number of adverse effects of this program on children's socio-emotional skills, possibly casting doubt on the use of one-sided *p*-values in this paper. The program they study is a warehousing childcare program, not a high quality early intervention program. Ramey and Ramey (2010) show that low quality childcare programs can have adverse effects. The Perry program was of extremely high quality and targeted highly disadvantaged children who generally lacked adequate parenting. Therefore, we should expect positive or no effects from the program.

[23] These are called mediation analyses in the statistics literature. See, e.g., Pearl (2012). Such analyses have been used for decades in economics. See, e.g., Klein and Goldberger (1955) and Theil (1958).

in measured skills on outcomes. We use econometric methods to address this problem. Using experimental variation, we can be more confident in the validity of our decompositions because we can test some of the assumptions maintained in our procedure. However, it is necessary to maintain some exogeneity assumptions in order to construct valid decompositions. This section makes those assumptions explicit.

Second, the Perry study has many highly correlated measurements of psychological skills that are laden with measurement error.[24] Moreover, the sample size of the study is small. We would exhaust the available degrees of freedom if we use all available psychological measurements to predict outcomes. Instead, we use factor analysis to create low dimensional, interpretable, and informative aggregates that summarize a range of psychological skills and account for measurement error.

Section IIA presents our model for outcomes. Section IIB presents our strategy for reducing numerous error-laden measurements to manageable summary measures and addressing the problem of measurement error. Section IIC discusses identification. We establish what features of the model are testable. Section IID summarizes a simple and robust three-step estimation procedure that is developed more extensively in the Appendix.

A. *The Outcome Equation*

Let D denote treatment assignment. $D = 1$ if an agent is treated and $D = 0$ otherwise. Let Y_1 and Y_0 be the counterfactual outcomes when D is fixed at "1" and "0" respectively. We use the subscript $d \in \{0, 1\}$ to represent variables when treatment is *fixed* at d. Fixing corresponds to manipulating treatment status d holding everything else constant.[25] The observed outcome is

$$Y = DY_1 + (1 - D)Y_0. \tag{1}$$

We assume that outcomes are independent across participants conditional on observed pre-program variables $\mathbf{X}$ that are assumed not to be affected by the program. We introduce the notion of skills that can be changed by the program and produce (in part) the treatment effect. The vector of skills when treatment is *fixed* at d is given by $\boldsymbol{\theta}_d = (\theta_d^j : j \in \mathcal{J})$, where $\mathcal{J}$ is an index set for skills. We define $\boldsymbol{\theta}$ in a fashion analogous to $Y : \boldsymbol{\theta} = D\boldsymbol{\theta}_1 + (1 - D)\boldsymbol{\theta}_0$.

Our analysis is based on the following linear model:

$$Y_d = \kappa_d + \boldsymbol{\alpha}_d \boldsymbol{\theta}_d + \boldsymbol{\beta}_d \mathbf{X} + \tilde{\epsilon}_d, \qquad d \in \{0, 1\}, \tag{2}$$

where κ_d is an intercept, $\boldsymbol{\alpha}_d$ and $\boldsymbol{\beta}_d$ are, respectively, $|\mathcal{J}|$-dimensional and $|\mathbf{X}|$-dimensional vectors of parameters where $|\mathbf{Q}|$ denotes the number of elements in $\mathbf{Q}$. While the pre-program variables $\mathbf{X}$ are assumed not to be affected by the

[24] For evidence on the extent of measurement error in these skills see Cunha and Heckman (2008) and Cunha, Heckman, and Schennach (2010).

[25] The distinction between fixing and conditioning traces back to Haavelmo (1943). See Pearl (2009) and Heckman and Pinto (2013) for recent discussions.

treatment, their effect on Y can be affected by the treatment. $\tilde{\epsilon}_d$ is a zero-mean error term assumed to be independent of regressors $\boldsymbol{\theta}_d$ and $\mathbf{X}$.

Perry analysts collected a rich array of measures of cognitive and personality skills. However, it is very likely that there are relevant skills that they did not measure. Notationally, let $\mathcal{J}_p \subseteq \mathcal{J}$ be the index set of skills on which we have measurements. The measurements may be imperfect so even these skills may not be directly observed. We decompose the term $\boldsymbol{\alpha}_d \boldsymbol{\theta}_d$ in equation (2) into components due to skills we measure and skills we do not:

$$
\begin{aligned}
(3) \quad Y_d &= \kappa_d + \sum_{j \in \mathcal{J}} \alpha_d^j \, \theta_d^j + \boldsymbol{\beta}_d \mathbf{X} + \tilde{\epsilon}_d \\
&= \kappa_d + \underbrace{\sum_{j \in \mathcal{J}_p} \alpha_d^j \, \theta_d^j}_{\substack{\text{skills on which} \\ \text{we have} \\ \text{measurements}}} + \underbrace{\sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha^j \, \theta_d^j}_{\substack{\text{skills on which} \\ \text{we have no} \\ \text{measurements}}} + \boldsymbol{\beta}_d \mathbf{X} + \tilde{\epsilon}_d \\
&= \tau_d + \sum_{j \in \mathcal{J}_p} \alpha_d^j \, \theta_d^j + \boldsymbol{\beta}_d \mathbf{X} + \epsilon_d,
\end{aligned}
$$

where $d \in \{0, 1\}$, $\tau_d = \kappa_d + \sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha_d^j \, E(\theta_d^j)$, and ϵ_d is a zero-mean error term defined by $\epsilon_d = \tilde{\epsilon}_d + \sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha_d^j (\theta_d^j - E(\theta_d^j))$. Any differences in the error terms between treatment and control groups can be attributed to differences in the skills on which we have no measurements. Without loss of generality we assume that $\tilde{\epsilon}_1 \overset{dist}{=} \tilde{\epsilon}_0$, where $\overset{dist}{=}$ means equality in distribution. Note that the error term ϵ_d is correlated with the measured skills if measured skills are correlated with unmeasured skills.

We seek to decompose treatment effects into components attributable to changes in the skills that we can measure. Assuming that changes in unmeasured skills attributable to the experiment are independent of $\mathbf{X}$, treatment effects can be decomposed into components due to changes in skills $E(\Delta\theta^j)$ and components due to changes in parameters $\Delta\alpha^j (= \alpha_1^j - \alpha_0^j)$:

$$
\begin{aligned}
(4) \quad E(\Delta Y_d | \mathbf{X}) &= E(Y_1 - Y_0 | \mathbf{X}) \\
&= (\tau_1 - \tau_0) + E\left(\sum_{j \in \mathcal{J}_p} \left(\alpha_1^j \, \theta_1^j - \alpha_0^j \, \theta_0^j \right) \right) + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)\mathbf{X} \\
&= (\tau_1 - \tau_0) \\
&\quad + E\left(\sum_{j \in \mathcal{J}_p} \left(\left(\Delta\alpha^j + \alpha_0^j \right) E(\Delta\theta^j) + (\Delta\alpha^j) \, E(\theta_0^j) \right) \right) \\
&\quad + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)\mathbf{X}.^{26}
\end{aligned}
$$

[26] Alternative decompositions are discussed in online Appendix E.3.

Equation (4) can be simplified if treatment affects skills, but not the impact of skills and background variables on outcomes, i.e., $\alpha_1^j = \alpha_0^j$; $j \in \mathcal{J}_p$ and $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0$.[27] Under the latter assumption, the term associated with $\mathbf{X}$ drops from the decomposition.

We establish below that if measured and unmeasured skills are independent in the no-treatment outcome equation, $\boldsymbol{\alpha}_0$ can be consistently estimated by a standard factor analysis. Under this assumption, and if $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_0$, we can test if the experimentally induced *increments* in unmeasured skills are independent of the experimentally induced *increments* in measured skills.[28] The intuition for this test is as follows. The skills for treated participants are the sum of the skills they would have had if they were assigned to the control group plus the increment due to treatment. If measured and unmeasured skill increments are independent, $\boldsymbol{\alpha}_1$ is consistently estimated by a standard factor analysis and we can test if plim $\hat{\boldsymbol{\alpha}}_1 =$ plim $\hat{\boldsymbol{\alpha}}_0$ where $(\hat{\boldsymbol{\alpha}}_0, \hat{\boldsymbol{\alpha}}_1)$ are estimates of $(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)$.[29] Assuming the exogeneity of $\mathbf{X}$, we can also test if plim $\hat{\boldsymbol{\beta}}_1 =$ plim $\hat{\boldsymbol{\beta}}_0$, where $(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1)$ are estimates of $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$. We test and do not reject these hypotheses.

Imposing these assumptions simplifies the notation. Equation (3) may be expressed as

$$Y_d = \tau_d + \sum_{j \in \mathcal{J}_p} \alpha^j \theta_d^j + \boldsymbol{\beta}\mathbf{X} + \epsilon_d, \qquad d \in \{0, 1\}. \tag{5}$$

In this notation, equation (1) becomes

$$\begin{aligned} Y &= D\underbrace{\left(\tau_1 + \sum_{j \in \mathcal{J}_p} \alpha^j \theta_1^j + \boldsymbol{\beta}\mathbf{X} + \epsilon_1\right)}_{Y_1} + (1 - D)\underbrace{\left(\tau_0 + \sum_{j \in \mathcal{J}_p} \alpha^j \theta_0^j + \boldsymbol{\beta}\mathbf{X} + \epsilon_0\right)}_{Y_0} \\ &= \tau_0 + \tau D + \sum_{j \in \mathcal{J}_p} \alpha^j \theta^j + \boldsymbol{\beta}\mathbf{X} + \epsilon, \end{aligned} \tag{6}$$

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 zero-mean error term, and $\theta^j = D\theta_1^j + (1 - D)\theta_0^j$, $j \in \mathcal{J}_p$ denote the skills that we can measure.

If the θ^j, $j \in \mathcal{J}_p$, are measured without error and are independent of the error term ϵ, least squares estimators of the parameters of equation (6) are unbiased for α^j, $j \in \mathcal{J}_p$.[30] If, on the other hand, the unmeasured skills are correlated with both measured skills and outcomes, least squares estimators of α^j, $j \in \mathcal{J}_p$, are biased and

[27] These are called structural invariance or autonomy assumptions in the econometric literature. See, e.g., Hurwicz (1962). These assumptions do not rule out heterogenous responses to treatment because $\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0$ may vary in the population.

[28] See online Appendix J for details.

[29] If skill increments are not independent, then in general even if $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_0$, plim $\boldsymbol{\alpha}_1 \neq$ plim $\boldsymbol{\alpha}_0$. Our test is valid in general even when $\boldsymbol{\alpha}_0$ cannot be consistently estimated. See online Appendix J. A distinct test of autonomy ($H_0: \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_0$) is possible if we maintain full exogeneity (i.e., measured skills are independent of unmeasured skills in both treatment regimes).

[30] Online Appendix G shows that the estimates of α in equation (6) are unbiased if measured and unmeasured skills are independent.

capture the effect of changes in the unmeasured skills as they are projected onto the measured components of $\boldsymbol{\theta}$, in addition to the direct effects of changes in measured components of $\boldsymbol{\theta}$ on Y.

Equation (6) is the basis for the decompositions reported in this paper. The treatment effect is

$$E(Y_1 - Y_0) = \underbrace{(\tau_1 - \tau_0)}_{\substack{\text{treatment effect due} \\ \text{to unmeasured skills}}} + \underbrace{\sum_{j \in \mathcal{J}_p} \alpha^j E(\theta_1^j - \theta_0^j)}_{\substack{\text{treatment effect due} \\ \text{to measured skills}}}. \tag{7}$$

Skill j can explain treatment effects only if it affects outcomes ($\alpha^j \neq 0$) and, on average, is affected by the experiment ($E(\theta_1^j - \theta_0^j) \neq 0$). We test both conditions.

Decomposition (7) would be straightforward to implement if the measured variables are independent of the unmeasured variables, and the measurements are accurate. In this case, the second term of (7) is easily constructed by using consistent estimates of the α^j and the effects of treatment on skills. However, psychological measurements are riddled with measurement error (Cunha and Heckman 2008). In addition, there are a large number of highly intercorrelated psychological measures that need to be condensed. We address these problems in this paper.

B. *Low-Dimensional Characterizations of Skills*

One way to summarize the psychological measures is to form simple unweighted indices constructed by taking averages of interpretable groups of items. This way of proceeding is widely used in psychology.[31] It is, however, fraught with difficulties. First, there are many ways to form aggregates. Second, the weightings of the measures used to form such aggregates are arbitrary. Third, this approach does not correct for measurement error, except through simple averaging.

This paper forms interpretable aggregates through factor analysis—a statistical method that summarizes the covariability among observed measures using low-dimensional latent variables. The method also accounts for measurement error.[32] We use the early measures of skills in the Perry data to extract the latent skills (θ^j; $j \in \mathcal{J}_p$) in equation (7) where the latent skills are the factors. We use a common measurement system for treated and untreated participants although θ_1^j, $j \in \mathcal{J}_p$ and θ_0^j, $j \in \mathcal{J}_p$, are allowed to differ.

More formally, let the index set for measures associated with factor $j \in \mathcal{J}_p$ be $\mathcal{M}^j$. Denote the measures for factor j in treatment group d by $M_{m^j,d}^j, d \in \{0, 1\}$. Henceforth, let $\boldsymbol{\theta}_d$ denote the vector of factors associated with the skills that can be measured in treatment state d, i.e., $\boldsymbol{\theta}_d = (\theta_d^j : j \in \mathcal{J}_p)$, $d \in \{0, 1\}$.

Following the psychometric literature summarized in Gorsuch (1983, 2003) and Thompson (2004), we assume that each measure is associated with at most one

[31] See the review in Borghans et al. (2008).

[32] See, e.g., Wansbeek and Meijer (2000). Table L.1 in online Appendix L gives estimates of the measurement error for the psychological measures used in this paper. In some measures, up to 80 percent of the variance is measurement error.

factor. We assume that the same measurement equations govern treatment and control groups so that the following equation is assumed to describe the relationship between the measures associated with factor j and the factor:

$$(8) \quad \text{Measures} : M^j_{m^j,d} = \nu^j_{m^j} + \varphi^j_{m^j}\theta^j_d + \eta^j_{m^j}, \quad j \in \mathcal{J}_p, \quad m^j \in \mathcal{M}^j.$$

To simplify the notation, we keep the covariates $\mathbf{X}$ implicit. Parameters $\nu^j_{m^j}$ are measure-specific intercepts. Parameters $\varphi^j_{m^j}$ are factor loadings. The ϵ_d in (5) and $\eta^j_{m^j}$ are mean-zero error terms assumed to be independent of $\boldsymbol{\theta}_d$, $d \in \{0, 1\}$, and of each other. The factor structure is characterized by the following equations:

$$(9) \quad \text{Factor Means} : E[\theta^j_d] = \mu^j_d, \qquad j \in \mathcal{J}_p$$

$$(10) \quad \text{Factor Covariance} : \text{Var}[\boldsymbol{\theta}_d] = \boldsymbol{\Sigma}_{\theta_d}, \qquad d \in \{0, 1\}.$$

The assumption that the parameters $\nu^j_{m^j}$, $\varphi^j_{m^j}$, $\text{Var}(n^j_{m^j}) : m^j \in \mathcal{M}^j, j \in \mathcal{J}_p$, do not depend on d simplifies the notation, as well as the interpretation of the estimates obtained from our procedure. It implies that the effect of treatment on the measured skills operates only through the latent skills and not through changing the measurement system for those skills. This assumption can be tested by estimating measurement systems separately for treatment and control groups and testing if measurement equation factor loadings and measurement equation intercepts differ between treatment and control groups.[33] We do not reject the hypotheses of equality of these parameters across treatment and control groups.[34]

C. *Identification*

Identification of factor models requires normalizations that set the location and scale of the factors (e.g., Anderson and Rubin 1956).[35] We set the location of each factor by fixing the intercepts of one measure—designated "the first"—to zero, i.e., $\nu^j_1 = 0$, $j \in \mathcal{J}_p$. This defines the location of factor j for each counterfactual condition. We set the scale of the factor by fixing the factor loadings of the first measure of each skill to one, i.e., $\varphi^j_1 = 1$, $j \in \mathcal{J}_p$. For all measures that are related to a factor (i.e., have a non-zero loading on the factor, $\varphi^j_{m^j}$), the decomposition of treatment effects presented in this paper is invariant to the choice of which measure is designated as the "first measure" for each factor and to any affine transformations of the measures.[36]

Identification is established in four steps. First, we identify the means of the factors, μ^j_d. Second, we identify the measurement factor loadings $\varphi^j_{m^j}$, the variances $\text{Var}(\eta^j_{m^j})$ of the measurement system, and the factor covariance structure $\boldsymbol{\Sigma}_{\theta_d}$.

[33] See online Appendix E.2 and the discussion in Section IIC.

[34] See online Appendix Tables L.2–L.4.

[35] We refer the reader to online Appendix E for a more detailed discussion of identification.

[36] See online Appendix E.3 for a proof. If $\boldsymbol{\alpha}_1 \neq \boldsymbol{\alpha}_0$, we acquire another term in the decomposition that is not invariant to affine transformations of the measures used to extract factors. However, even in this case, the treatment effect arising from measured skills in (7) is invariant. See also Heckman and Pinto (forthcoming) for a more general analysis of the combinations of parameters identified under monotonic transformations of the measures.

Third, we use the parameters identified from the first and second steps to secure identification of the measurement intercepts $\nu^j_{m^j}$. Finally, we use the parameters identified in the first three steps to identify the factor loadings $\boldsymbol{\alpha} = (\alpha^j; j \in \mathcal{J}_p)$ and intercept τ_d of the outcome equations. We discuss each of these steps in turn.

Factor Means.—We identify μ^j_1 and μ^j_0 from the mean of the designated first measure for treatment and control groups: $E(M^j_{1,d}) = \mu^j_d, j \in \mathcal{J}_p, d \in \{0, 1\}$.

Measurement Loadings.—From the covariance structure of the measurement system, we can identify: (i) the factor loadings of the measurement system $\varphi^j_{m^j}$; (ii) the variances of the measurement error terms, $\text{Var}(\eta^j_{m^j})$; and (iii) the factor covariance matrix, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_d}$. Factors are freely correlated. We need at least three measures for each skill $j \in \mathcal{J}_p$, all with non-zero factor loadings.[37] The $\varphi^j_{m^j}$ can depend on $d \in \{0, 1\}$, and we can identify $\varphi^j_{m^j,d}$. We test if $H_0: \varphi^j_{m^j,1} = \varphi^j_{m^j,0}, j \in \mathcal{J}_p$, and we do not reject these hypotheses.[38]

Measurement Intercepts.—From the means of the measurements, i.e., $E(M^j_{m^j,d}) = \nu^j_{m^j} + \varphi^j_{m^j}\, \mu^j_d$, we identify $\nu^j_{m^j}$, $m^j \in \mathcal{M}^j \backslash \{1\}$, $j \in \mathcal{J}_p$. Recall that the factor loadings $\varphi^j_{m^j}$ and factor means μ^j_d are identified. Assuming equality of the intercepts $(\nu^j_{m^j})$ between treatment and control groups guarantees that treatment effects on measures, i.e., $E(M^j_{m^j,1}) - E(M^j_{m^j,0})$, operate solely through treatment effects on factor means, i.e., $\mu^j_1 - \mu^j_0$. However, identification of our decomposition requires intercept equality only for the designated first measure of each factor. We test and do not reject $H_0: \nu^j_{m^j,1} = \nu^j_{m^j,0}$ for all $m^j \in \mathcal{M}^j \backslash \{1\}, j \in \mathcal{J}_p$.[39]

Outcome Equation.—Adult outcome factor loadings in equation (5) can be identified using the covariances between outcomes and the designated first measure of each skill. We form the covariances of each outcome Y_d with the designated first measure of each skill $j \in \mathcal{J}_p$ to obtain $\text{Cov}(Y_d, \mathbf{M}_{1,d}) = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_d}\, \boldsymbol{\alpha}$ where $\boldsymbol{\alpha} = (\alpha^j; j \in \mathcal{J}_p)$. By the previous argument, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_d}$ is identified. Thus $\boldsymbol{\alpha}$ is identified whenever $\det(\boldsymbol{\Sigma}_{\boldsymbol{\theta}_d}) \neq 0$. Outcome factor loadings $\boldsymbol{\alpha}$ can depend on $d \in \{0, 1\}$, as they can be identified through $\text{Cov}(Y_d, \mathbf{M}_{1,d}) = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_d}\, \boldsymbol{\alpha}_d$, which can be separately identified for treatments and controls. We test $H_0: \alpha^j_1 = \alpha^j_0, j \in \mathcal{J}_p$, and we do not reject these hypotheses.[40] Using $E(Y_d)$, we can identify τ_d because all of the other parameters of each outcome equation are identified.

[37] Having three measures allows us to form three covariances and to solve for the three unrestricted parameters of the three-measurement system. With two measures, we form one covariance which cannot, by itself, be used to identify the two unrestricted parameters of the two measurement system. See Anderson and Rubin (1956).

[38] Table L.4 in the online Appendix. Proof of identification of this more general model is given in online Appendix E.

[39] See Table L.4 in the online Appendix.

[40] Tables L.2 and L.3 in the online Appendix.

D. *Estimation Procedure*

We estimate the model using a simple three stage procedure. First, we estimate the measurement system. Second, from these equations we can estimate the skills for each participant. Third, we estimate the relationship between participant skills and lifetime outcomes. Proceeding in this fashion makes identification and estimation transparent. In Section IV and online Appendix L we show that a one-step procedure produces estimates very similar to those obtained from the three-step procedure. We estimate the model separately for males and females in light of the evidence that there are strong gender differences in program effects.[41]

We compute p-values using the bootstrap. We draw $K = 1{,}000$ bootstrap samples of the original data and apply the estimation procedure to each pseudo-sample the same way we apply it to the original data. For a one-tailed test with an upper tail rejection region, the bootstrap p-value is estimated by

$$\hat{p}(\hat{\varrho}) = \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}(\varrho_k^* > \hat{\varrho}), \tag{11}$$

where $\hat{\varrho}$ is the parameter of interest as estimated from the original data, and ϱ_k^* is the k-th draw from the bootstrap data-generating process satisfying the null hypothesis.[42] We describe the details of our estimation procedure in the Appendix.

III. Measures of Cognitive and Personality Skills

This section explains how we condense the data on psychological skills. Using standard psychometric methods, we establish that only three factors are required to explain the available psychological measures. The extracted factors have clear interpretations. After extracting the factors, we test the validity of the derived system.

A. *Our Measure of Cognition*

A large literature establishes the importance of cognition, as measured by IQ, in explaining a variety of life outcomes (see, e.g., Gottfredson 1997, and Jensen 1998, for surveys). We use the Stanford-Binet Intelligence Test (Terman and Merrill 1960) as our measure of cognition. Mean differences in Stanford-Binet scores between treatment and control groups are plotted by age and gender in panels A (for males) and B (for females) of Figure 1. A boost in the IQs of children in the treatment group is observed soon after the program starts at age three. A few years after the program ends, the effect of treatment on IQ essentially disappears for males. A small, borderline statistically significant, positive effect remains for females. In our analysis, we use IQs at ages seven, eight, and nine, since this is the period when the

[41] See Heckman et al. (2010a) and Tables L.5 and L.6 in online Appendix L.
[42] See Wasserman (2006) for details.

treatment effect on IQ becomes relatively stable for both genders, and IQ becomes rank stable after those ages.[43]

B. *Personality Skills*

The most influential taxonomy of personality skills is the Big Five personality inventory (McCrae and John 1992; John and Srivastava 1999).[44] Unfortunately, the Big Five was developed long after the Perry experiment was conducted. We only have access to psychological measures of personality skills collected before the Big Five was codified.[45]

Perry Measures of Personality Skills.—There are 43 child personality measures in the Perry data. These measures belong to two separate psychological inventories of personality skills: the Pupil Behavior Inventory (PBI) and the Ypsilanti Rating Scale (YRS). These measures are displayed in Tables C.1 and D.1 in the online Appendix, where their correspondence with the Big Five inventory is noted.[46] The PBI inventory was developed by Vinter et al. (1966) to measure behavioral and attitudinal factors that affect academic success. The YRS measures were developed by the Perry analysts to measure academic potential and socio-emotional skills (Weikart, Bond, and McNeil 1978). The PBI and YRS questionnaires consist of multiple questions called "items." They were given to teachers in classes attended by Perry students after the program was completed to assess students in their classes. An example of one of the PBI items is "lying or cheating," and the possible answers are (1) very frequently, (2) frequently, (3) sometimes, (4) infrequently, and (5) very infrequently. The YRS questionnaire asks questions about socio-emotional skills, such as academic potential and social development.[47]

C. *Approaches to Summarizing the Data*

There are many ways to summarize the available psychological measures in an interpretable fashion. One way is to form indices of measures using the groupings employed by the Perry psychologists.[48] The PBI and YRS scales were designed to alert educators of behavioral and motivational problems of children in school. There is considerable overlap among items in the groupings, and the relationship of these measures to more interpretable psychological constructs is unclear.

[43] Measures of IQ from alternative tests, the Leiter and the PPVT, evolve in a fashion similar to scores on the Stanford-Binet test (see panels A–D of Figure B.6 of online Appendix B). Among IQ measures available in the Perry data, Stanford-Binet is the most established one. The Stanford-Binet IQ test is widely used and has high reliability (see Santrock 2008). Online Appendix B presents a detailed description of the test. Borghans et al. (2008) and Almlund et al. (2011) discuss the evidence on the rank stability of IQ after age ten or so.

[44] However, even the Big Five is not universally accepted. See Borghans et al. (2008), and Almlund et al. (2011), for surveys of personality psychology literature.

[45] The skills can be related to the Big Five (see Tables C.1 and D.1 of the online Appendix).

[46] We thank Angela Duckworth for helping us make this correspondence.

[47] Although measures are missing on the PBI and YRS, the longitudinal structure of the Perry experiment allows us to use closely adjacent measures by age to impute the missing data. See online Appendices C and D.

[48] There are nine such groupings. We list the corresponding measures within groups in Tables C.1 and D.1 of the online Appendix.

An intuitively appealing way to construct summary measures is to ignore the groupings used by the Perry psychologists and to select measures from all of their scales based on common sense and previous research in psychology, such as their interpretability in terms of the Big Five. This procedure is called "operationalization" in psychology, and is inherently subjective.[49] There are many ways to form such indices, which leads to a complex model selection problem.

The approach used in this paper is to apply exploratory factor analysis (EFA) to the available measures. EFA is a standard statistical method widely used in psychometrics to allocate measures to factors (e.g., Gorsuch 2003; Thompson 2004). It is used to form the Big Five (see Goldberg 1993). EFA establishes links between a small number of latent factors and the available measures. Each measure is allowed to depend on at most one factor, and the derived factors are allowed to be freely correlated.

Our application of EFA produces three interpretable factors which we interpret as cognition, externalizing behavior, and academic motivation.[50] The derived factors are consistent with previous research in psychology on the predictive power of psychological skills. We implement EFA in two stages. We first select the number of factors. We then allocate the measures of personality skills to different factors.

D. *Exploratory Factor Analysis*

We use several accepted procedures to determine the number of factors: the scree test (Cattell 1966), Onatski's test (2009), and Horn's (1965) parallel analysis test.[51] Overall, these procedures point to a three-factor characterization for both men and women.[52]

Exploratory Factor Analysis: Establishing Dedicated Measures.—We extract factors following the criteria for EFA laid out by Gorsuch (1983).[53] This method is widely used although its application requires judgment on the part of the analyst (see Gorsuch 2003; and Thompson 2004). EFA identifies blocks of measures that are strongly correlated within each block (i.e., satisfy *convergent validity*), but are weakly correlated between blocks (i.e., satisfy *discriminant validity*). It discards measures that load on multiple factors. Application of standard EFA methodology to the 46 cognitive and personality Perry measures gives the 13 measures displayed in Table H.2 in the online Appendix.[54]

[49] See Borghans et al. (2008) and Almlund et al. (2011).

[50] The interpretation of a factor is derived from the interpretation of the measures from which it is extracted.

[51] Online Appendix H provides additional details about the criteria used in the literature to choose the number of factors. The Guttman-Kaiser rule (Guttman 1954; and Kaiser 1960, 1961) suggests 7–9 factors. This rule is well known to overestimate the number of factors (see Zwick and Velicer 1986; Gorsuch 2003, and Thompson 2004) and this makes it less informative than the other methods. We do not use it in this paper.

[52] See Table H.1 in the online Appendix.

[53] See online Appendix H.1 for the algebra of factor rotation and for the definition of various factor selection criteria. We use *direct quartimin* oblique rotation. We find that other widely recognized oblique rotation methods such as *geomin*, lead to similar results and the same choice of measures as quartimin (see Table L.7 of the online Appendix, which shows the same pattern as the direct quartimin solution in Table H.2 of online Appendix H). This result is in line with the literature showing that widely recognized methods of oblique rotation produce similar results (Fabrigar et al. 1999).

[54] Measures are retained if they are strongly related to one and only one factor. For statements about the standard EFA methodology, see Tabachnick and Fidell (2001), Thompson (2004), and Costello and Osborne (2005).

Table 2—Cognitive and Personality Factors and Their Measures

Cognition		Externalizing behavior		Academic motivation	
Measures[a]	Age	Measures[a]	Age[b]	Measures[a]	Age[b]
Stanford-Binet IQ	7	Disrupts classroom procedures	7–9	Shows initiative	7–9
Stanford-Binet IQ	8	Swears or uses obscene words	7–9	Alert and interested in school work	7–9
Stanford-Binet IQ	9	Steals	7–9	Hesitant to try, or gives up easily	7–9
		Lying or cheating	7–9		
		Influences others toward troublemaking	7–9		
		Aggressive toward peers	7–9		
		Teases or provokes students	7–9		
Cronbach's alpha,[c] males	0.838	Cronbach's alpha, males	0.906	Cronbach's alpha, males	0.901
Cronbach's alpha, females	0.913	Cronbach's alpha, females	0.916	Cronbach's alpha, females	0.896

Notes: [a]See online Appendix B for a detailed description of Stanford-Binet IQ measures. Externalizing behavior and academic motivation are proxied by items of the Pupil Behavior Inventory (PBI) described in online Appendix C. PBI items are described in this table the same way they appear in the questionnaire. For example, "lying or cheating" and "steals" were the full descriptions of misbehavior that teachers were asked to evaluate. [b]Each personality measure is an average over nonmissing observations at ages seven, eight, and nine. [c]Cronbach's alpha (see Cronbach 1951) is a statistic that captures how well a set of measures proxies a latent skill. Cronbach's alpha is the lower bound of the internal consistency reliability of measures that are proxies for a skill. The internal consistency reliability is defined as the square of the correlation between the measured scale defined as the sum of the measures and the underlying skill θ (Allen and Yen 2002). In our case, the correlations between the skills and the scales (equal to the square roots of alphas) range from 0.70 to 0.82 for males and from 0.80 to 0.84 for females. The Cronbach's alphas can also be interpreted as a correlation between the observed scale and a hypothetical alternative scale measuring the same skill and based on the same number of hypothetical alternative items (Nunnally and Bernstein 1994). For this table, the alphas (i.e., the correlations) range from 0.84 to 0.91 for males and from 0.90 to 0.92 for females.

We follow Gorsuch (2003) and Thompson (2004), and derive a fully dedicated system as described by equations (8)–(10), i.e., a system in which each measure is associated with at most one factor. This procedure is called confirmatory factor analysis (CFA), which Gorsuch (2003) and Thompson (2004) advocate as the next step after conducting EFA. It produces the interpretable system displayed in Table 2, based on three factors: cognition, externalizing behavior, and academic motivation.[55] Externalizing behavior is proxied by measures of behavior related to lying, stealing, and swearing, as well as being aggressive and disruptive. It has been linked to crime and aggressive behavior.[56] Academic motivation is proxied by measures of student interest, persistence, and initiative in learning and is linked to performance in schools as measured by achievement tests. The personality measures proxying externalizing behavior and academic motivation are conceptually related to Big Five factors in personality psychology.[57]

The Predictive Power of Externalizing Behavior and Academic Motivation.—The factors extracted by the EFA procedure are closely linked to skills that have been

[55] The factor loadings for the dedicated system are presented in Table L.8 of online Appendix L.

[56] Externalizing behavior is linked to the Big Five measures of agreeableness, neuroticism, and conscientiousness. See Almlund et al. (2011).

[57] See Almlund et al. (2011) for a discussion of these relationships.

shown to be predictive of adult outcomes. This gives us greater confidence in using them to explain the Perry treatment effects.

The recent literature in economics shows that externalizing behavior predicts child and adult outcomes (Segal 2008, forthcoming).[58] The literature in psychology shows that externalizing behavior is negatively associated with academic achievement.[59] Childhood externalizing behaviors have also been shown to be related to adolescent and adult delinquency (e.g., Nagin and Tremblay 1999; and Broidy et al. 2003).

The literature in criminology and psychology demonstrates that early antisocial behaviors are highly predictive of adolescent and adult antisocial behaviors (Huesmann et al. 1984; Olweus 1979; Gersten et al. 1976). Antisocial behaviors measured between ages seven and eleven strongly predict criminal behaviors in adulthood (Moffitt 1993; Loeber 1982). Meanwhile, disobedient and aggressive behaviors measured as early as ages three to five predict later childhood conduct disorders and adolescent arrests (Moffitt 1993; White et al. 1990). Most children with conduct disorders experience social difficulties in adulthood, whereas only 8 percent of children without conduct disorders experience such difficulties (Zoccolillo et al. 1992). Similarly, many children with antisocial behavior around ages eight to ten become antisocial adults (Robins 1978; Coie et al. 1995; Olweus 1979)[60] and chronic criminal offenders (Loeber 1982). Almost all antisocial adults were antisocial children (Robins 1978). Our analysis confirms previous evidence on the stability of antisocial skills into adulthood. We find stable rank correlations between externalizing behavior at ages seven to nine and subsequent measures of crime as late as ages 19, 27, and 40 (see Figure 4). The evidence from the literature in psychology and criminology joined with the evidence from this paper suggests that reducing early externalizing behavior reduces crime.

Academic motivation, apart from its obvious link to performance in school, has been shown to be a statistically significant predictor of decreased drug use (Bryant et al. 2003, and Razzino et al. 2004) and alcohol consumption (Zimmerman and Schmeelk-Cone 2003; Simons-Morton 2004; and Vaughan, Corbin, and Fromme 2009). Since drinking and drug use are associated with crime (Anglin and Perrochet 1998; and Greenfeld 1998), youth with higher levels of academic motivation are less likely to engage in criminal activities. Flouri and Buchanan (2002) show that for both males and females, low academic motivation in adolescence is positively related to trouble with the police at age 16. Cymbalisty, Schuck, and Dubeck (1975) show that for males who have already committed crimes, recidivism decreases with motivation for learning. Therefore, it is expected that experimentally induced enhancements in academic motivation would be a source of treatment effects for education and crime outcomes. We confirm such effects for education, but not for crime.

[58] For more information about externalizing behaviors, see Achenbach (1978), Campbell et al. (1996), and Brunnekreef et al. (2007).

[59] See Richman, Stevenson, and Graham (1982); Egeland et al. (1990); Jimerson, Egeland, and Teo (1999); and Jimerson et al. (2002).

[60] Robins (1978) estimates that 36–41 percent of children with antisocial behaviors become highly antisocial adults.

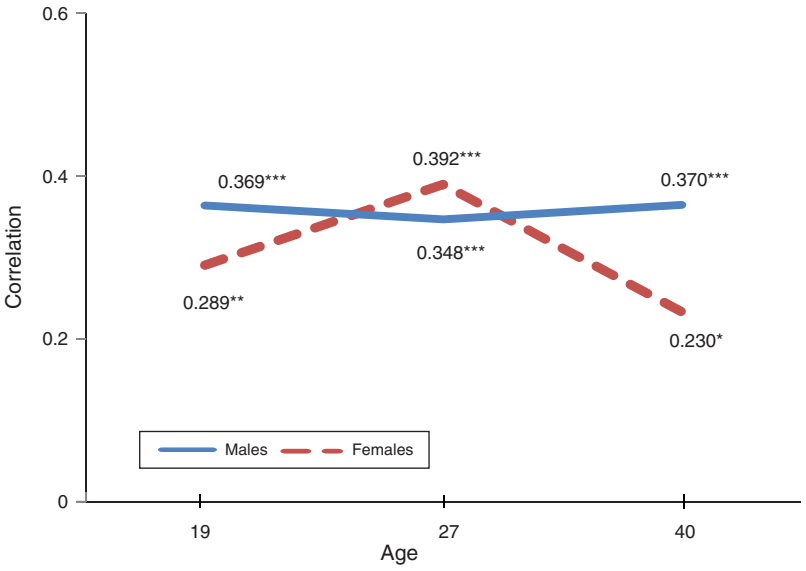

FIGURE 4. SPEARMAN'S RANK CORRELATIONS BETWEEN EXTERNALIZING BEHAVIOR AT AGES 7–9 AND NUMBER OF ARRESTS BY AGES 19, 27, AND 40

Note: One-sided tests.
***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

IV. The Effect of Treatment on Skills and the Sources of Treatment Effects

We first study how treatment affects the extracted factors. We then investigate how the factors affect life outcomes. Finally, we decompose adult treatment effects into components corresponding to changes in each factor. The first analysis is based on the output of step two of the three-step estimation procedure described in detail in the Appendix. The second and third analyses are based on the output of the third step of the procedure.

A. *The Effect of the Perry Program on Cognitive and Personality Skills*

Figure 5 graphs kernel densities of factor scores and presents one-sided *p*-values for testing the equality of the means for each skill between the treatment and control groups.[61] The Perry program has a statistically significant treatment effect on externalizing behavior at the 5 percent level for males and at the 1 percent level for females. The effects on cognition and academic motivation are statistically significant at the 10 percent and 5 percent levels, respectively, for females, but are not statistically significant for males. This evidence is consistent with the evidence in Table 1 of a statistically significant treatment effect on achievement test scores, which is much stronger for girls than for boys.

[61] Cognition is uncorrelated with externalizing behavior, while academic motivation correlates with both externalizing behavior and cognition (see Table L.9 of online Appendix L).

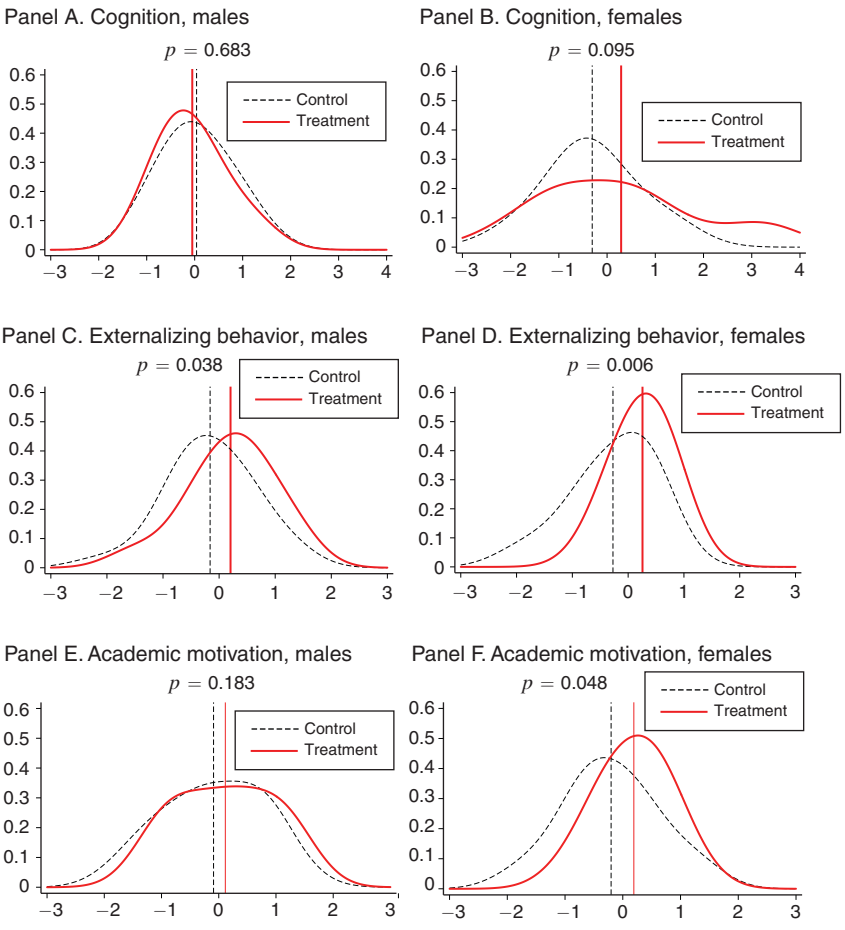

FIGURE 5. KERNEL DENSITIES OF FACTOR SCORES

Notes: Probability density functions of Bartlett (1937) factor scores are shown. Densities are computed based on a normal kernel. Numbers above the charts are one-sided p-values testing the equality of factor score means for the treatment and control groups. Higher externalizing behavior corresponds to more socially desirable behavior. See online Appendix L for the empirical CDFs of the factor scores (Figure L.5). Vertical lines locate factor score means for treatment and control groups.

The kernel densities reveal different patterns of the effect of the program on the distribution of skills. The cognition of females is enhanced mostly in the right tail of the distribution (panel B). In contrast, a substantial part of the improvement in externalizing behavior for females operates through enhancing low levels of the skill (panel D). Externalizing behavior in males is improved at all levels. Academic motivation in females is improved at all levels except for the top percentiles (see panel F). There is no statistically significant difference in the distribution of cognition for males (panel A).[62]

[62] We also test for gender differences in skills and find that differences are not statistically significant. In other words, for each skill and for each treatment group we cannot reject the null hypothesis of equality of skills between males and females. See Figure L.1 of online Appendix L.

B. *The Effect of Cognitive and Personality Skills on Outcomes*

In order to estimate the effects of factors on outcomes we estimate a model described by a system of equations (5) and (8) conditioning on background variables $\mathbf{X}$. We present estimates of $\boldsymbol{\alpha} = (\alpha^j, j \in \mathcal{J}_p)$ in equation (7). Table 3 shows that all three factors (cognition, externalizing behavior, and academic motivation) have statistically significant effects on at least one outcome.

Different factors affect different outcomes. Cognition primarily affects achievement tests and also affects certain labor market outcomes. Externalizing behavior affects crime, labor market outcomes, and health behaviors. Academic motivation boosts educational outcomes and reduces long-term unemployment.

Treatment effects are generated through changes in skills if (i) skills affect outcomes and (ii) skills are enhanced by the intervention. Thus, even though cognition and academic motivation are positively related to the CAT scores of males, the absence of a relationship for males between treatment and both cognition and academic motivation makes this channel either weak or nonexistent.

C. *Decomposing Treatment Effects on Outcomes by Source*

Figures 6 and 7 present our estimated decompositions of treatment effects into experimentally induced improvements in cognition, externalizing behavior, academic motivation, and other factors. By "other factors" we mean the residual treatment effect associated with unmeasured skills. We report the percentage of each treatment effect attributable to each component.[63] The numbers shown above each component are one-sided *p*-values for the test of whether the component is zero. We stress that these decompositions are invariant to the choice of normalizing measures and to affine transformations of the measures.[64]

We decompose the treatment effect for a number of outcomes: performance on the California Achievement Test (CAT), special education at school and mental impairment, labor market outcomes such as income and employment, health behaviors such as smoking tobacco or using drugs, marriage duration, and crime outcomes. The crime outcomes are especially important since they are the dominant component of Perry program's total benefit (e.g., Heckman et al. 2010b). We only decompose treatment effects that have been shown to be statistically significant at the 10 percent level or below after adjusting for the effects of multiple-hypothesis testing on significance levels (Heckman et al. 2010a; Conti et al. 2013). Proceeding in this fashion leads to somewhat different decompositions for males and females.

The effect of the intervention on life outcomes operates primarily through the program's enhancement of externalizing behavior. Components attributable to changes in this factor are generally statistically significant and, in most cases, explain

[63] Figures 6 and 7 are slightly simplified representations of the results presented in Tables L.10 and L.11 of the online Appendix. To simplify the exposition, for the figures in the main text, contributions opposite in sign to those of the total treatment effect are set to zero. These contributions are small and statistically insignificant. Thus the figures in the text are an accurate summary of the essential information in the tables. Appendix L explains the methodology for constructing the figures and shows (see Figure L.2) that the figures presented in the text closely approximate the actual decomposition.

[64] See online Appendix E.

TABLE 3—FACTOR LOADINGS OF OUTCOME EQUATIONS

Outcome	Cognition		Externalizing behavior		Academic motivation		Sample size
	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value	
Panel A. Males							
CAT total at age 14, end of grade 8 (+)	0.819***	(0.000)	−0.203	(0.845)	0.700***	(0.000)	45
Number of misdemeanor arrests, age 27 (−)	−0.259	(0.359)	−1.226**	(0.028)	−0.152	(0.367)	59
Number of felony arrests, age 27 (−)	−0.618	(0.235)	−1.333**	(0.023)	0.219	(0.557)	59
Number of adult arrests (misd.+fel.), age 27 (−)	−0.876	(0.251)	−2.559**	(0.014)	0.067	(0.549)	59
Monthly income, age 27 (+)	0.970**	(0.038)	0.698**	(0.046)	−0.257	(0.670)	55
Use tobacco, age 27 (−)	−0.179	(0.121)	−0.332***	(0.001)	0.159	(0.847)	57
Number of misdemeanor arrests, age 40 (−)	−0.620	(0.383)	−2.424*	(0.087)	0.196	(0.501)	59
Number of felony arrests, age 40 (−)	−0.628	(0.266)	−1.755**	(0.014)	0.293	(0.570)	59
Number of adult arrests (misd.+fel.), age 40 (−)	−1.248	(0.327)	−4.180**	(0.039)	0.489	(0.525)	59
Number of lifetime arrests, age 40 (−)	−1.100	(0.359)	−4.740**	(0.030)	0.239	(0.519)	59
Employed, age 40 (+)	0.277**	(0.012)	0.230**	(0.011)	−0.270	(0.991)	54
Panel B. Females							
CAT total, age 8 (+)	0.219**	(0.039)	−0.134	(0.729)	0.689***	(0.000)	35
CAT total, age 14 (+)	0.154	(0.113)	−0.448	(0.931)	0.899***	(0.001)	31
Any special education, age 14 (−)	−0.041	(0.273)	0.119	(0.759)	−0.209*	(0.064)	37
Mentally impaired at least once, age 19 (−)	−0.039	(0.283)	0.227	(0.948)	−0.308***	(0.008)	33
Number of misdemeanor violent crimes, age 27 (−)	0.083	(0.778)	−1.080**	(0.043)	0.150	(0.700)	37
Number of felony arrests, age 27 (−)	0.021	(0.609)	−0.451*	(0.053)	0.140	(0.808)	37
Jobless for more than 1 year, age 27 (−)	0.139	(0.920)	0.048	(0.608)	−0.465***	(0.003)	36
Ever tried drugs other than alcohol or weed, age 27 (−)	−0.043	(0.201)	−0.146	(0.144)	0.122	(0.854)	34
Number of misdemeanor violent crimes, age 40 (−)	0.084	(0.774)	−1.078**	(0.043)	0.081	(0.592)	37
Number of felony arrests, age 40 (−)	0.047	(0.704)	−0.589**	(0.014)	0.078	(0.643)	37
Number of lifetime violent crimes, age 40 (−)	0.096	(0.807)	−1.220**	(0.023)	0.165	(0.704)	37
Months in all marriages, age 40 (+)	21.748	(0.111)	13.591	(0.289)	10.453	(0.280)	36

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 normalized to control mean zero and variance of one. See Tables L.14 and L.15 of online Appendix L for more detailed versions of this table containing coefficients for background variables. Monthly income is adjusted to thousands of year-2006 dollars using annual national CPI.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

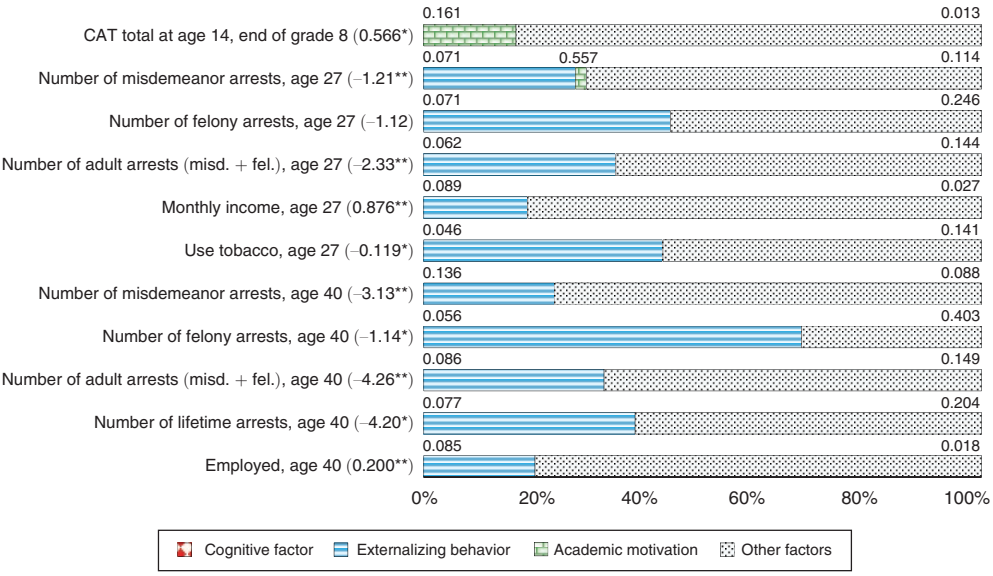

FIGURE 6. DECOMPOSITIONS OF TREATMENT EFFECTS ON OUTCOMES, MALES

Notes: 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. The figure is a slightly simplified visualization of online Appendix Tables L.10 and L.14: small and statistically insignificant contributions of the opposite sign are set to zero. See online Appendix L for detailed information about the simplifications made to produce the figure. "CAT total" denotes California Achievement Test total score normalized to control mean zero and variance of one. Monthly income is adjusted to thousands of year-2006 dollars using annual national CPI.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

20–60 percent of the treatment effects on crime for males and about 40–60 percent for females (see Figures 6 and 7).

The mediating effects of externalizing behavior are not only statistically significant, but also economically significant. Reported arrests and registered crimes are only a small fraction of the actual number of crimes. For instance, only one in 15 property crimes and one in five violent crimes actually leads to an arrest.[65] We find that experimentally induced reductions in externalizing behavior (by one standard deviation) lead to a decline in the total number of lifetime arrests by statistically significant 1.7 ($p = 0.077$) and the number of felony arrests by 0.6 ($p = 0.056$) for males at age 40.[66] For females, the total number of felony arrests by age 40 is reduced by 0.31 ($p = 0.050$), and the number of registered lifetime violent crimes is reduced by 0.65 ($p = 0.046$).[67] The reduction in actual crimes

[65] Heckman et al. (2010b) estimate that the average victimization to arrest ratio in Midwestern urban areas is 15.0 for property crimes and 5.3 for violent crimes.

[66] Control group means for the number of total lifetime and felony arrests for males are 12.4 and 3.2, with standard errors 1.9 and 0.7.

[67] Tables L.10 and L.11 of the online Appendix present the effects in terms of absolute levels rather than in relative levels as shown in Figures 6 and 7. Control group means for the number of lifetime felony arrests and number of registered lifetime violent crimes are 0.42 and 0.65 respectively, with standard errors 0.18 and 0.29.

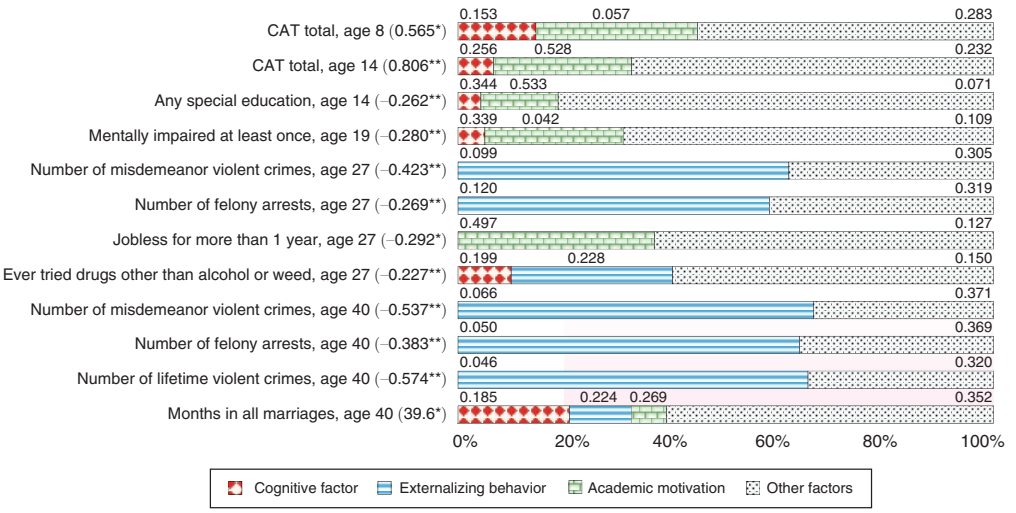

FIGURE 7. DECOMPOSITIONS OF TREATMENT EFFECTS ON OUTCOMES, FEMALES

Notes: 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 in each outcome. The figure is a slightly simplified visualization of online Appendix Tables L.11 and L.15: small and statistically insignificant contributions of the opposite sign are set to zero. See online Appendix L for detailed information about the simplifications made to produce the figure. "CAT total" denotes California Achievement Test total score normalized to control mean zero and variance of one.

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

is likely several times larger than these reductions in the number of arrests and registered crimes. Since externalizing behavior is both malleable at early ages (see Figure 5) and strongly predictive of crime (see Table 3), it should not be surprising that crime reduction has been found to be a major benefit of the Perry program.

We also decompose the effect of the program on an achievement test (CAT) for both males and females. For females, enhancements in academic motivation explain about 30 percent of the treatment effect on CAT scores at age eight. This estimate is statistically significant at a 10 percent level ($p = 0.057$). For CAT scores at age 14, the role of academic motivation is not precisely determined for males or for females ($p = 0.161$ and 0.528).

Finally, we decompose a number of education, labor market, and health outcomes. Academic motivation consistently explains a share of treatment effects for all education-related outcomes, which is not surprising given strong links between academic motivation and education outcomes presented in Table 3. However, only some components of these decompositions are precisely determined (e.g., CAT and the status of being mentally impaired for females).

For labor market outcomes, we find that about 20 percent of the treatment effect on monthly income at age 27 ($p = 0.089$) and also about 20 percent of the treatment effect on the probability of employment at age 40 ($p = 0.085$) are explained by early improvements in externalizing behavior. Additionally, externalizing behavior explains about 40 percent of tobacco use at age 27 ($p = 0.046$).

D. *Decompositions Based on Indices versus Decompositions Using Factor Scores*

A simple alternative to our factor procedure for summarizing the evidence on the effects of experimentally induced changes in measurements on outcomes is to use indices of the measures in place of factor scores. Such indices, which are unweighted averages over the measures, are commonly used. In making this comparison we take as given the cluster of measures identified by application of EFA (see Table 2). Picking the clusters in this fashion avoids the serious practical problem that many groupings are possible, and many tests among competing specifications are nonnested. The comparison being made is one between using unweighted averages of measures not correcting for measurement error and a method that extracts factors by weighting the measures by the estimated factor loadings and adjusting for measurement error.

Figure 8 decomposes a selection of the treatment effects for a variety of outcomes for each gender using both indices and factor models.[68] For each outcome, we show the results of the two different estimation procedures using a pair of bars. The first bar in each pair corresponds to the estimates from the procedure used in this paper. These bars are identical to those presented in Figures 6 and 7. The second bar in each pair corresponds to the decomposition obtained from estimating equation (7) using indices.

Comparing the first bar with the second bar reveals that even though results of the two procedures lead to similar qualitative conclusions about the role of mediating skills, the estimates of the explained treatment effect components and the associated p-values are numerically different. As is apparent from equation (7), there are two possible sources of difference in the decompositions: (i) different factor loadings α^j and (ii) different estimates of the treatment effect on the factors: $E(\theta_1^j - \theta_0^j)$, $j \in \mathcal{J}_p$. The approach using indices only partially corrects for attenuation bias by reducing measurement error through simple averaging. Our factor approach explicitly addresses measurement error. Thus the index approach likely generates downward-biased decompositions. Indeed, the shares of the treatment effects explained by externalizing behavior are generally smaller for the index-based procedure. The p-values from the index-based procedure are somewhat smaller for females but somewhat larger for males. Most of the comparisons presented in Figure 8 are consistent with this interpretation.[69, 70]

[68] We present a full comparison of indices with factors in online Appendix L for all treatment effects analyzed in this paper. See Figures L.6 and L.7 to be compared to Figures 6 and 7 respectively.

[69] Upward bias may arise for the following reason. An index uses equal weights for all measures. It may happen that for some particular outcome, measures that are more predictive of that outcome have relatively higher weights in the index compared to their weights on the true factor scores. Then the index will be more predictive of that particular outcome than the factor score. If the effect of using an index outweighs the attenuation bias due to measurement error, it may happen that a decomposition based on an index is biased upwards. There are only two instances of this occurring in Figure 8 (tobacco use and felony arrests for males).

[70] A related issue, somewhat tangential to the main point of this paper, is the decomposition that results from using achievement test scores in place of IQ test scores in constructing our decomposition. As noted by Borghans et al. (2011a, b), 30–40 percent of the variation in achievement test scores is attributable to variation in personality scores. Thus the common practice of assuming that achievement tests proxy intelligence is misleading. When we use the CAT scores at ages seven to nine in place of IQ scores at those ages, and construct an "achievement" cognition factor, we overstate the importance of "cognition" as a source of treatment effects. See Tables L.12 and L.13, Figure L.3, and the discussion in online Appendix L.

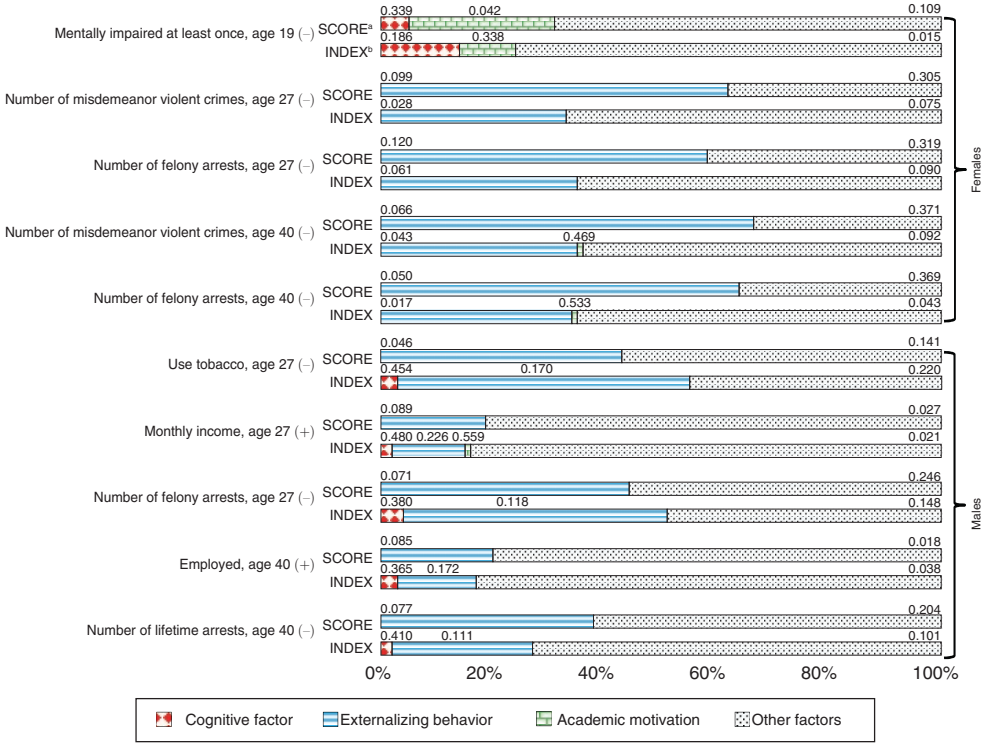

FIGURE 8. DECOMPOSITIONS OF TREATMENT EFFECTS BY INDICES VERSUS FACTOR SCORES

Notes: The total treatment effect is normalized to 100 percent. One-sided *p*-values are shown above each component in each outcome. The figure is a slightly simplified visualization of results from Tables L.10, L.11, K.1, and K.2 of the online Appendix; small and statistically insignificant contributions of the opposite sign are set to zero. See online Appendix L for detailed information about the simplifications used to produce the figure.

[a]"SCORE" denotes models where personality skills are measured by factor scores.

[b]"INDEX" denotes models where personality skills are measured by indices constructed using unweighted averages over the items.

E. *More Efficient Estimates*

The three-step estimation procedure used in this paper is simple and intuitive. In general, it is not statistically efficient given that we do not impose cross-equation restrictions across the stages of the estimation. In online Appendix L, we compare the decompositions obtained from our three-step estimation procedure to those obtained from a one-step maximum likelihood estimation method where the measurement system and outcome equation are estimated jointly.[71] The results from both procedures are in close agreement, although *p*-values from the maximum likelihood procedure are generally lower.[72]

[71] See online Appendix L, Figure L.4.

[72] Tables L.5 and L.6 of online Appendix L show the full set of estimates for the decompositions.

F. *Testing the Validity of the Derived System*

The procedure used to create the dedicated factor system is based in part on judgments by the analyst. Such judgments are widely used in the psychometric literature.[73] To gain greater confidence in the system created by our EFA analysis, we test the validity of the derived factor structure.

Our application of the EFA methodology yields 13 dedicated measures out of 46 available cognitive and personality measures. The 33 unused measures do not survive the EFA testing criteria. To test the validity of this specification, we run a series of tests on the measurement and outcome equations. We first determine if, conditional on the extracted factors, the unused measures exhibit a treatment effect. If they do not, they are not candidates for explaining the treatment effect for outcomes. We also determine whether, conditional on the extracted factors, the unused measures explain outcomes. Evidence from both types of tests support the low-dimensional specification of equations (5) and (8)–(10) derived from applying EFA. Online Appendix M presents a detailed discussion of these tests.

G. *A Framework for Unifying Diverse Studies of Child Development*

The framework developed in this paper facilitates the interpretation of diverse treatment effects within and across programs as the manifestations of program-induced changes in a low-dimensional set of skills of participants. This framework can be used to unify the interpretation of the treatment effects across different studies with different interventions applied to different populations. By focusing on the channels through which the different programs produce their effects, we gain a deeper understanding of the skills that matter and how they can be affected by various influences on the child. Systematic application of this framework will enable the intervention literature to move beyond meta-analyses to understand the common mechanisms producing success in children and how different interventions boost different skills to different degrees. This framework also offers a basis for unifying observational studies of family influence with intervention studies.[74] Investments made by families boost $\boldsymbol{\theta}$ as do the investments made in intervention programs. Using the framework developed and applied in this paper, we can in principle compare family investments and interventions in terms of their effects on $\boldsymbol{\theta}$.

V. Conclusions

Using experimental data from an influential early childhood program, we analyze the sources of program treatment effects. Coupling experimental variation with an econometric model, we estimate the role of enhancements in cognition, externalizing behavior, and academic motivation in producing the Perry treatment effects. Persistent changes in personality skills play a substantial role in producing the success of the Perry program. The reduction in externalizing behavior, which explains

[73] See Gorsuch (2003) and Thompson (2004).
[74] For a template of this research program see Cunha and Heckman (2009).

the bulk of the effects of the Perry program on criminal, labor market, and health behavior outcomes, is especially strong.[75]

We offer a new understanding of how a few hours per day of preschool at ages three and four with a curriculum that promotes social competency, planning, and organization can significantly and beneficially affect life outcomes. The importance and malleability of these skills deserves greater emphasis in public policies designed to promote skills and alleviate poverty.[76]

Appendix on the Three-Step Estimation Procedure

Step 1: For a given set of dedicated measurements, and choice of the number of factors, we estimate the factor model using measurement system (8)–(10).

Step 2: We use the measures and factor loadings estimated in the first step to compute a vector of *factor scores* for each participant i. We form unbiased estimates of the true vector of skills $\boldsymbol{\theta}_i = (\theta_i^j; j \in \mathcal{J}_p)$ for agent i. The factor measurement equations contain $\mathbf{X}$ which we suppress to simplify the expressions. Notationally, we represent the measurement system for agent i as

$$\underbrace{\mathbf{M}_i}_{|\mathcal{M}|\times 1} = \underbrace{\boldsymbol{\varphi}}_{|\mathcal{M}|\times|\mathcal{J}_p|}\ \underbrace{\boldsymbol{\theta}_i}_{|\mathcal{J}_p|\times 1} + \underbrace{\boldsymbol{\eta}_i}_{|\mathcal{M}|\times 1}, \tag{A1}$$

where $\boldsymbol{\varphi}$ represents a matrix of the factor loadings estimated in the first step and $\mathbf{M}_i$ is the vector of stacked measures for participant i subtracting the intercepts ν_{mj}^j of equation (8). The dimension of each element in equation (A1) is shown beneath it, where $\mathcal{M} = \cup_{j\in\mathcal{J}_p} \mathcal{M}^j$ is the union of all the index sets of the measures $\mathcal{M}^j$, $j \in \mathcal{J}_p$. The error term for agent i, $\boldsymbol{\eta}_i$, has zero mean and is independent of the vector of skills $\boldsymbol{\theta}_i$. $\text{Cov}(\boldsymbol{\eta}_i, \boldsymbol{\eta}_i) = \boldsymbol{\Omega}$. The most commonly used estimator of factor scores is based on a linear function of measures: $\boldsymbol{\theta}_{S,i} = \mathbf{L}'\mathbf{M}_i$. Unbiasedness requires that $\mathbf{L}'\boldsymbol{\varphi} = \mathbf{I}_{|p|}$, where $\mathbf{I}_{|p|}$ is a $|p|$-dimensional identity matrix.[77] To achieve unbiasedness, $\mathbf{L}$ must satisfy $\mathbf{L}' = (\boldsymbol{\varphi}'\boldsymbol{\Omega}^{-1}\boldsymbol{\varphi})^{-1}\boldsymbol{\varphi}'\boldsymbol{\Omega}^{-1}$. The unbiased estimator of the factor is

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

Factor score estimates can be interpreted as the output of a *GLS* estimation procedure where measures are taken as dependent variables and factor loadings are

[75] Our analysis cannot rule out the possibility that the initial enhancement of IQ in the Perry program permanently boosted personality skills (e.g., by giving participants more understanding of their environment and promoting self-confidence and other skills) even though the initial IQ surge faded. To examine this possibility would require estimating the state space model of Cunha and Heckman (2008) to examine the transient dynamics of the model. This would be a formidable empirical challenge for a sample the size of the Perry study.

[76] Since we analyze one program in one site with one level of program intensity, we are unable to determine the external validity of our evidence for other sites or intensity levels, nor can we discuss how easy it is to go to scale with the program. An analysis of these questions for early childhood outcomes is possible using data from ICPSR (2010) because roughly 30 percent of Head Start centers adopt some version of the Perry curriculum.

[77] The method is due to Bartlett (1937) and is based on the restricted minimization of mean squared error, subject to $\mathbf{L}'\boldsymbol{\varphi} = \mathbf{I}_{|\mathcal{J}|}$.

treated as regressors. By the Gauss-Markov theorem, for a known φ the proposed estimator is the best linear unbiased estimator of the vector of skills $\boldsymbol{\theta}_i$.[78]

Step 3: The use of factor scores instead of the true factors to estimate equation (5) generates biased estimates of outcome coefficients $\boldsymbol{\alpha}$. Even though estimates of $\boldsymbol{\theta}_i$ are unbiased, there is still a discrepancy between the true and measured $\boldsymbol{\theta}_i$ due to estimation error. To correct for the bias, we implement a bias-correction procedure. Because we estimate the variance of $\boldsymbol{\theta}$ and the variance of the measurement errors in the first step of our procedure, we can eliminate the bias created by the measurement error.

Consider the outcome model for agent i:

$$Y_i = \boldsymbol{\alpha}\boldsymbol{\theta}_i + \boldsymbol{\gamma}\mathbf{Z}_i + \boldsymbol{\epsilon}_i, \tag{A2}$$

where $(\boldsymbol{\theta}_i, \mathbf{Z}_i) \perp\!\!\!\perp \boldsymbol{\epsilon}_i$ and $E(\boldsymbol{\epsilon}_i) = 0$. For brevity of notation, we use $\mathbf{Z}_i$ to denote pre-program variables, treatment status indicators, and the intercept term of equation (5). From equation (A1), the factor scores $\boldsymbol{\theta}_{S,i}$ can be written as the skills $\boldsymbol{\theta}_i$ plus a measurement error $\mathbf{V}_i$; that is,

$$\boldsymbol{\theta}_{S,i} = \boldsymbol{\theta}_i + \mathbf{V}_i \text{ such that } (\mathbf{Z}_i, \boldsymbol{\theta}_i) \perp\!\!\!\perp \mathbf{V}_i \text{ and } E(\mathbf{V}_i) = 0. \tag{A3}$$

Replacing $\boldsymbol{\theta}_i$ with $\boldsymbol{\theta}_{S,i}$ yields $Y_i = \boldsymbol{\alpha}\boldsymbol{\theta}_{S,i} + \boldsymbol{\gamma}\mathbf{Z}_i + \boldsymbol{\epsilon}_i - \boldsymbol{\alpha}\mathbf{V}_i$. The linear regression estimator of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ is inconsistent:

$$\text{plim}\begin{pmatrix}\hat{\boldsymbol{\alpha}}\\ \hat{\boldsymbol{\gamma}}\end{pmatrix} = \underbrace{\begin{pmatrix}\text{Cov}(\boldsymbol{\theta}_S, \boldsymbol{\theta}_S) & \text{Cov}(\boldsymbol{\theta}_S, \mathbf{Z})\\ \text{Cov}(\mathbf{Z}, \boldsymbol{\theta}_S) & \text{Cov}(\mathbf{Z}, \mathbf{Z})\end{pmatrix}^{-1}\begin{pmatrix}\text{Cov}(\boldsymbol{\theta}, \boldsymbol{\theta}) & \text{Cov}(\boldsymbol{\theta}, \mathbf{Z})\\ \text{Cov}(\mathbf{Z}, \boldsymbol{\theta}) & \text{Cov}(\mathbf{Z}, \mathbf{Z})\end{pmatrix}}_{\mathbf{A}}\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\gamma}\end{pmatrix}. \tag{A4}$$

This is the multivariate version of the standard one-variable attenuation bias formula. All covariances in $\mathbf{A}$ can be computed directly except for the terms that involve $\boldsymbol{\theta}$. The covariance $\text{Cov}(\boldsymbol{\theta}, \boldsymbol{\theta})$ is estimated in step 1. Using equation (A3), we can compute $\text{Cov}(\mathbf{Z}, \boldsymbol{\theta}_S) = \text{Cov}(\mathbf{Z}, \boldsymbol{\theta})$. Thus, $\mathbf{A}$ is identified. Our bias-correction procedure consists of pre-multiplying the least squares estimators $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$ by $\mathbf{A}^{-1}$, thus providing consistent estimates of $(\boldsymbol{\alpha}, \boldsymbol{\gamma})$.[79] A one step maximum likelihood procedure, while less intuitive, directly estimates the parameters without constructing the factors and accounts for measurement error. It is justified in large samples under standard regularity conditions. It produces estimates very close to those obtained from the three-step procedure but with smaller standard errors. See online Appendix L.

[78] Online Appendix F discusses other estimators considered in the literature. Note that the assumption that φ is known can be replaced with the assumption that φ is consistently estimated and we can use an asymptomatic version of the Gauss-Markov theorem replacing "unbiased" with "unbiased in large samples."

[79] See Croon (2002) for more details on this bias-correction approach.

REFERENCES

Achenbach, Thomas M. 1978. "The Child Behavior Profile: I. Boys Aged 6–11." *Journal of Consulting and Clinical Psychology in the Schools* 46 (3): 478–88.

Allen, Mary J., and Wendy M. Yen. 2002. *Introduction to Measurement Theory*. Prospect Heights, IL: Waveland Press.

Almlund, Mathilde, Angela Lee Duckworth, James J. Heckman, and Tim D. Kautz. 2011. "Personality Psychology and Economics." In *Handbook of the Economics of Education*. Vol. 4, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 1–181. Amsterdam: Elsevier, North-Holland.

Almond, Douglas, and Janet Currie. 2011. "Human Capital Development Before Age Five." *In Handbook of Labor Economics.* Vol. 4B, edited by Orley Ashenfelter and David Card, 1315–1486. Amsterdam: Elsevier, North-Holland.

Anderson, Michael L. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association* 103 (484): 1481–95.

Anderson, Theodore W., and Herman Rubin. 1956. "Statistical Inference in Factor Analysis." In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 5, edited by Jerzy Neyman, 111–50. Berkeley: University of California Press.

Anglin, M. Douglas, and Brian Perrochet. 1998. "Drug Use and Crime: A Historical Review of Research Conducted by the UCLA Drug Abuse Research Center." *Substance Use and Misuse* 33 (9): 478–88.

Baker, Michael, Jonathan Gruber, and Kevin Milligan. 2008. "Universal Child Care, Maternal Labor Supply, and Family Well-Being." *Journal of Political Economy* 116 (4): 709–45.

Barnett, W. Steven. 1995. "Long-Term Effects of Early Childhood Programs on Cognitive and School Outcomes." *The Future of Children* 5 (3): 25–50.

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

Belfield, Clive R., Milagros Nores, Steve Barnett, and Lawrence Schweinhart. 2006. "The High/Scope Perry Preschool Program: Cost-Benefit Analysis Using Data from the Age-40 Followup." *Journal of Human Resources* 41 (1): 162–90.

Berrueta-Clement, John R., Lawrence J. Schweinhart, Steven W. Barnett, Ann S. Epstein, and David P. Weikart. 1984. *Changed Lives: The Effects of the Perry Preschool Program on Youths Through Age* 19. Ypsilanti, MI: High/Scope Foundation.

Bodrova, Elena, and Deborah J. Leong. 2007. *Tools of the Mind: The Vygotskian Approach to Early Childhood Education*. Upper Saddle River, NJ: Pearson/Merrill Prentice Hall.

Borghans, Lex, Angela Lee 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. 2011a. "Identification Problems in Personality Psychology." *Personality and Individual Differences* 51 (3): 315–20.

Borghans, Lex, Bart H. H. Golsteyn, James J. Heckman, and John Eric Humphries. 2011b. "IQ, Achievement, and Personality." Unpublished.

Bowles, Samuel, and Herbert Gintis. 1976. *Schooling in Capitalist America: Educational Reform and the Contradictions of Economic Life*. New York: Basic Books.

Bowles, Samuel, and Herbert Gintis. 2001. "Inheritance of Economic Status: Education, Class, and Genetics." In *International Encyclopedia of the Social and Behavioral Sciences*, edited by Neil J. Smelser and Paul B. Baltes, 4132–41. Oxford, UK: Pergamon.

Broidy, Lisa M., Daniel S. Nagin, Richard E. Tremblay, John E. Bates, Bobby Brame, Kenneth A. Dodge, David Fergusson, et al. 2003. "Developmental Trajectories of Childhood Disruptive Behaviors and Adolescent Delinquency: A Six-Site, Cross-National Study." *Developmental Psychology* 39 (2): 222–45.

Brunnekreef, J. Agnes, Leo M. J. De Sonneville, Monika Althaus, Ruud B. Minderaa, Albertine J. Oldehinkel, Frank C. Verhulst, and Johan Ormel. 2007. "Information Processing Profiles of Internalizing and Externalizing Behavior Problems: Evidence from a Population-based Sample of Preadolescents." *Journal of Child Psychology and Psychiatry* 48 (2): 185–93.

Bryant, Alison L., John E. Schulenberg, Patrick M. O'Malley, Jerald G. Bachman, and Lloyd D. Johnston. 2003. "How Academic Achievement, Attitudes, and Behaviors Relate to the Course of Substance Use During Adolescence: A 6-Year, Multiwave National Longitudinal Study." *Journal of Research on Adolescence* 13 (3): 361–97.

Campbell, Frances A., and Craig T. Ramey. 1994. "Effects of Early Intervention on Intellectual and Academic Achievement: A Follow-Up Study of Children from Low-Income Families." *Child Development* 65 (2): 684–98.

Campbell, Frances A., and Craig T. Ramey. 1995. "Cognitive and School Outcomes for High-Risk African-American Students at Middle Adolescence: Positive Effects of Early Intervention." *American Educational Research Journal* 32 (4): 743–72.

Campbell, Susan B., Elizabeth W. Pierce, Ginger Moore, Susan Marakovitz, and Kristin Newby. 1996. "Boys' Externalizing Problems at Elementary School Age: Pathways from Early Behavior Problems, Maternal Control, and Family Stress." *Development and Psychopathology* 8 (4): 701–19.

Carneiro, Pedro, and James J. Heckman. 2003. "Human Capital Policy." In *Inequality in America: What Role for Human Capital Policies?*, edited by James J. Heckman, Alan B. Krueger and Benjamin M. Friedman, 77–239. Cambridge, MA: MIT Press.

Cattell, Raymond B. 1966. "The Scree Test for the Number of Factors." *Multivariate Behavioral Research* 1 (2): 245–76.

Coie, John, Robert Terry, Kari Lenox, John Lochman, and Clarine Hyman. 1995. "Childhood Peer Rejection and Aggression as Predictors of Stable Patterns of Adolescent Disorder." *Development and Psychopathology* 7 (4): 697–13.

Conti, Gabriella, James Heckman, Seong Moon, and Rodrigo Pinto. 2013. "Long-term Health Effects of Early Childhood Interventions." Unpublished.

Costello, Anna B., and Jason W. Osborne. 2005. "Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis." *Practical Assessment, Research and Evaluation* 10 (7): 173–78.

Cronbach, Lee J. 1951. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika* 16 (3): 297–334.

Croon, Marcel A. 2002. "Using Predicted Latent Scores in General Latent Structure Models." In *Latent Variable and Latent Structure Models*, edited by George A. Marcoulides and Irini Moustaki, 195–224. Mahwah, 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): 738–82.

Cunha, Flavio, and James J. Heckman. 2009. "The Economics and Psychology of Inequality and Human Development." *Journal of the European Economic Association* 7 (2–3): 320–64.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78 (3): 883–931.

Cymbalisty, Bohdan Y., Solomon Z. Schuck, and John A. Dubeck. 1975. "Achievement Level, Institutional Adjustment and Recidivism among Juvenile Delinquents." *Journal of Community Psychology* 3 (3): 289–94.

DeLong, J. Bradford, and Konstantin Magin. 2009. "The US Equity Return Premium: Past, Present, and Future." *Journal of Economic Perspectives* 23 (1): 193–208.

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

Diamond, Adele, Steven Barnett, Jessica Thomas, and Sarah Munro. 2007. "Preschool Program Improves Cognitive Control." *Science* 318 (5855): 1387–88.

Egeland, Byron, Mark Kalkoske, Natan Gottesman, and Martha F. Erickson. 1990. "Preschool Behavior Problems: Stability and Factors Accounting for Change." *Journal of Child Psychology and Psychiatry* 31 (6): 891–909.

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 (3): 272–99.

Flouri, Eirini, and Ann Buchanan. 2002. "Father Involvement in Childhood and Trouble With the Police in Adolescence: Findings From the 1958 British Cohort." *Journal of Interpersonal Violence* 17 (6): 689–701.

Gersten, Joanne C., Thomas S. Langner, Jeanne G. Eisenberg, Ora Simcha-Fagan, and Elizabeth D. McCarthy. 1976. "Stability and Change in Types of Behavioral Disturbance of Children and Adolescents." *Journal of Abnormal Child Psychology* 4 (2): 111–27.

Goldberg, Lewis R. 1993. "The Structure of Phenotypic Personality Traits." *American Psychologist* 48 (1): 26–34.

Gorsuch, Richard. 1983. *Factor Analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gorsuch, Richard L. 2003. "Factor Analysis." In *Handbook of Psychology: Research Methods in Psychology*. Vol. 2, edited by Irving B. Weiner, John A. Schinka, and Wayne F Velicer, 143–64. Hoboken, NJ: John Wiley and Sons, Inc.

Gottfredson, Linda S. 1997. "Why g Matters: The Complexity of Everyday Life." *Intelligence* 24 (1): 79–132.

Gray, Susan W., and Rupert A. Klaus. 1970. "The Early Training Project: A Seventh-Year Report." *Child Development* 41 (4): 909–24.

Greenfeld, Lawrence A. 1998. "Alcohol and Crime: An Analysis of National Data on the Prevalence of Alcohol Involvement in Crime." US Department of Justice, Bureau of Justice Statistics, National Criminal Justice Reference Service, no. 168632.

Guttman, Louis. 1954. "Some Necessary Conditions for Common-factor Analysis." *Psychometrika* 19 (2): 149–61.

Haavelmo, Trygve. 1943. "The Statistical Implications of a System of Simultaneous Equations." *Econometrica* 11 (1): 1–12.

Hanushek, Eric, and Alfred A. Lindseth. 2009. *Schoolhouses, Courthouses, and Statehouses: Solving the Funding-Achievement Puzzle in America's Public Schools*. Princeton, NJ: Princeton University Press.

Hanushek, Eric A., and Ludger Woessmann. 2008. "The Role of Cognitive Skills in Economic Development." *Journal of Economic Literature* 46 (3): 607–68.

Heckman, James J. 2000. "Policies to Foster Human Capital." *Research in Economics* 54 (1): 3–56.

Heckman, James J. 2008. "Schools, Skills, and Synapses." *Economic Inquiry* 46 (3): 289–324.

Heckman, James J., and Tim Kautz. 2012. "Hard Evidence on Soft Skills." *Labour Economics* 19 (4): 451–64.

Heckman, James, Seong Hyeok Moon, Rodrigo Pinto, Peter Savelyev, and Adam Yavitz. 2010a. "Analyzing Social Experiments as Implemented: A Reexamination of the Evidence from the HighScope Perry Preschool Program." *Quantitative Economics* 1 (1): 1–46.

Heckman, James J., Seong Hyeok Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam Yavitz. 2010b. "The Rate of Return to the HighScope Perry Preschool Program." *Journal of Public Economics* 94 (1–2): 114–28.

Heckman, James J., and Rodrigo Pinto. Forthcoming. "Econometric Mediation Analyses: Identifying the Sources of Treatment Effects from Experimentally Estimated Production Technologies with Unmeasured and Mismeasured Inputs." *Econometric Reviews*.

Heckman, James J., and Rodrigo Pinto. 2013. "Causal Analysis After Haavelmo." Unpublished.

Heckman, James, Rodrigo Pinto, and Peter Savelyev. 2013. "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes: Dataset." *American Economic Review*. http://dx.doi.org/10.1257/aer.103.6.2052.

Heckman, James J., Jora Stixrud, and Sergio Urzúa. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24 (3): 411–82.

Herrnstein, Richard J., and Charles A. Murray. 1994. *The Bell Curve: Intelligence and Class Structure in American Life*. New York: Free Press.

Horn, John L. 1965. "A Rationale and Test for the Number of Factors in Factor Analysis." *Psychometrika* 30 (2): 179–85.

Huesmann, L. Rowel, Leonard D. Eron, Monroe M. Lefkowitz, and Leopold O. Walder. 1984. "Stability of Aggression Over Time and Generations." *Developmental Psychology* 20 (6): 1120–34.

Hurwicz, Leonid. 1962. "On the Structural Form of Interdependent Systems." In *Logic, Methodology and Philosophy of Science*, edited by E. Nagel, P. Suppes, and A. Tarski, 232–39. Palo Alto, CA: Stanford University Press.

Inter-University Consortium for Political and Social Research (ICPSR). 2010. *Head Start Impact Study, 2002–2006: Spring 2003 Center Director Interview Data Codebook*. Ann Arbor, MI: ICPSR.

Jensen, Arthur R. 1998. *The g Factor: The Science of Mental Ability*. Westport, CT: Praeger.

Jimerson, Shane, Byron Egeland, and Adrian Teo. 1999. "A Longitudinal Study of Achievement Trajectories: Factors Associated with Change." *Journal of Educational Psychology* 91 (1): 116–26.

Jimerson, Shane R., Phillip Ferguson, Angela D. Whipple, Gabrielle E. Anderson, and Michael J. Dalton. 2002. "Exploring the Association Between Grade Retention and Dropout: A Longitudinal Study Examining Socio-Emotional, Behavioral, and Achievement Characteristics of Retained Students." *The California School Psychologist* 7: 51–62.

John, Oliver P., and Sanjay Srivastava. 1999. "The Big Five Trait Taxonomy: History, Measurement and Theoretical Perspectives." In *Handbook of Personality: Theory and Research*, edited by Lawrence A. Pervin and Oliver P. John, 102–38. New York: The Guilford Press.

Kaiser, Henry F. 1960. "The Application of Electronic Computers to Factor Analysis." *Educational and Psychological Measurement* 20 (1): 141–51.

Kaiser, Henry F. 1961. "A Note on Guttman's Lower Bound for the Number of Common Factors." *British Journal of Statistical Psychology* 14 (1): 1–2.

Klein, Lawrence Robert, and Arthur Stanley Goldberger. 1955. *An Econometric Model of the United States*, 1929–1952. Amsterdam: North-Holland Publishing Company.

Knudsen, Eric I., James J. Heckman, Judy L. Cameron, and Jack P. Shonkoff. 2006. "Economic, Neurobiological and Behavioral Perspectives on Building America's Future Workforce." *Proceedings of the National Academy of Sciences* 103 (27): 10155–62.

Lazar, Irving, Richard Darlington, Harry Murray, Jacqueline Royce, Ann Snipper, and Craig T. Ramey. 1982. "Lasting Effects of Early Education: A Report from the Consortium for Longitudinal Studies." *Monographs of the Society for Research in Child Development* 47 (2/3): 1–151.

Loeber, Rolf. 1982. "The Stability of Antisocial and Delinquent Child Behavior: A Review." *Child Development* 53 (6): 1431–46.

McCrae, Robert R., and Oliver P. John. 1992. "An Introduction to the Five-factor Model and Its Applications." *Journal of Personality* 60 (2): 175–215.

Moffitt, Terrie E. 1993. "Adolescence-Limited and Life-Course-Persistent Antisocial Behavior: A Developmental Taxonomy." *Psychological Review* 100 (4): 674–701.

Nagin, Daniel, and Richard E. Tremblay. 1999. "Trajectories of Boys' Physical Aggression, Opposition, and Hyperactivity on the Path to Physically Violent and Nonviolent Juvenile Delinquency." *Child Development* 70 (5): 1181–96.

Nunnally, Jum C., and Ira H. Bernstein. 1994. *Psychometric Theory*. New York: McGraw-Hill.

Obama, Barack. 2013. "Fact Sheet: President Obama's Plan for Early Education for all Americans." The White House, Office of the Press Secretary. Washington, DC. http://www.whitehouse.gov/the-press-office/2013/02/13/fact-sheet-president-obama-s-plan-early-education-all-americans.

Olweus, Dan. 1979. "Stability of Aggressive Reaction Patterns in Males: A Review." *Psychological Bulletin* 86 (4): 852–75.

Onatski, Alexei. 2009. "Testing Hypotheses about the Number of Factors in Large Factor Models." *Econometrica* 77 (5): 1447–79.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*. 2nd ed. New York: Cambridge University Press.

Pearl, Judea. 2012. "The Mediation Formula: A Guide to the Assessment of Causal Pathways in Nonlinear Models." In *Causality: Statistical Perspectives and Applications*, edited by Carlo Berzuini, Philip Dawid, and Luisa Bernardinelli, 151–79. Hoboken, NJ: John Wiley and Sons, Inc.

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

Ramey, Craig T., and Sharon Landesman Ramey. 2010. "Children's Right to Thrive: The Foundational Years." Unpublished.

Razzino, Brian E., Sheila C. Ribordy, Kathryn Grant, Joseph R. Ferrari, Blake S. Bowden, and Jennifer Zeisz. 2004. "Gender-related Processes and Drug Use: Self-expression with Parents, Peer Group Selection, and Achievement Motivation." *Adolescence* 39 (153): 167–77.

Reynolds, Arthur J., Judy A. Temple, Dylan L. Robertson, and Emily A. Mann. 2001. "Long-term Effects of an Early Childhood Intervention on Educational Achievement and Juvenile Arrest: A 15-Year Follow-up of Low-Income Children in Public Schools." *Journal of the American Medical Association* 285 (18): 2339–46.

Richman, Naomi, James E. Stevenson, and Philip Jeremy Graham. 1982. *Preschool to School: A Behavioural Study*. San Diego: Academic Press.

Robins, Lee N. 1978. "Sturdy Childhood Predictors of Adult Antisocial Behaviour: Replications from Longitudinal Studies." *Psychological Medicine* 8 (4): 611–22.

Santrock, John W. 2008. *A Topical Approach to Life-Span Development*. 4th ed. New York: McGraw-Hill.

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

Schweinhart, Lawrence J., Jeanne Montie, Zongping Xiang, W. Steven Barnett, Clive R. Belfield, and Milagros Nores. 2005. *Lifetime Effects: The High/Scope Perry Preschool Study Through Age* 40. Ypsilanti, MI: High/Scope Press.

Schweinhart, Lawrence J., and David P. Weikart. 1981. "Effects of the Perry Preschool Program on Youths Through Age 15." *Journal of Early Intervention* 4 (1): 29–39.

Segal, Carmit. Forthcoming. "Misbehavior, Education, and Labor Market Outcomes." *Journal of the European Economic Association.*

Segal, Carmit. 2008. "Classroom Behavior." *Journal of Human Resources* 43 (4): 783–814.

Simons-Morton, Bruce. 2004. "Prospective Association of Peer Influence, School Engagement, Drinking Expectancies, and Parent Expectations with Drinking Initiation among Sixth Graders." *Addictive Behaviors* 29 (2): 299–309.

Sylva, Kathy. 1997. "The Quest for Quality in Curriculum." An appended commentary to *Lasting Differences: The High/Scope Preschool Curriculum Comparison Study through Age* 23, by Lawrence J. Schweinhart and David P. Weikart. Ypsilanti, MI: High/Scope Press.

Tabachnick, Barbara G., and Linda S. Fidell. 2001. *Using Multivariate Statistics*. 4th ed. Boston: Allyn and Bacon.

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

Theil, Henri. 1958. *Economic Forecasts and Policy*. Amsterdam: North-Holland Publishing Company.

Thompson, **Bruce.** 2004. *Exploratory and Confirmatory Factor Analysis : Understanding Concepts and Applications*. Washington, DC: American Psychological Association.

Tough, **Paul.** 2009. "Can the Right Kinds of Play Teach Self-Control?" *New York Times*, September 25.

Vaughan, **Ellen L.**, **William R. Corbin**, **and Kim Fromme.** 2009. "Academic and Social Motives and Drinking Behavior." *Psychology of Addictive Behaviors* 23 (4): 564–76.

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*. Edited by Alex Kozulin. Cambridge, MA: MIT Press.

Wansbeek, **Tom**, **and Erik Meijer.** 2000. *Measurement Error and Latent Variables in Econometrics*. *Advanced Textbooks in Economics*. Volume 37. Amsterdam: Elsevier Science, North-Holland.

Wasserman, **Larry.** 2006. *All of Nonparametric Statistics*. Springer Texts in Statistics. New York: Springer.

Weikart, **David P.** 1967. "Preliminary Results From a Longitudinal Study of Disadvantaged Preschool Children." Education Resources Information Center (ERIC) No. 030 490. Paper presented at the Convention of the Council for Exceptional Children, St. Louis, MO.

Weikart, **David P.**, **James T. Bond**, **and Judy T. McNeil.** 1978. *The Ypsilanti Perry Preschool Project: Preschool Years and Longitudinal Results Through Fourth Grade*. Monograph of the High/Scope Educational Research Foundation, No. 3. Ypsilanti, MI: High/Scope Press.

White, **Jennifer L.**, **Terrie E. Moffitt**, **Felton Earls**, **Lee Robins**, **and Phil A. Silva.** 1990. "How Early Can We Tell?: Predicators of Childhood Conduct Disorder and Adolescent Delinquency." *Criminology* 28 (4): 507–33.

Yoshikawa, **Hirokazu.** 1995. "Long-Term Effects of Early Childhood Programs on Social Outcomes and Delinquency." *The Future of Children* 5 (3): 51–75.

Zimmerman, **Marc A.**, **and Karen H. Schmeelk-Cone.** 2003. "A Longitudinal Analysis of Adolescent Substance Use and School Motivation among African American Youth." *Journal of Research on Adolescence* 13 (2): 185–210.

Zoccolillo, **Mark**, **Andrew Pickles**, **David Quinton**, **and Michael Rutter.** 1992. "The Outcome of Childhood Conduct Disorder: Implications for Defining Adult Personality Disorder and Conduct Disorder." *Psychological Medicine* 22 (4): 971–86.

Zwick, **William R.**, **and Wayne F. Velicer.** 1986. "Comparison of Five Rules for Determining the Number of Components to Retain." *Psychological Bulletin* 99 (3): 432–42.