Uncovering Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects *

Michael Anderson
MIT Department of Economics

April 2006

Abstract

The view that the returns to public educational investments are highest for early childhood interventions stems primarily from several influential randomized trials - Abecedarian, Perry, and the Early Training Project - that point to super-normal returns to preschool interventions. This paper implements a unified statistical framework to present a de novo analysis of these experiments, focusing on core issues that have received little attention in previous analyses: treatment effect heterogeneity by gender, over-rejection of the null hypothesis due to multiple inference, and robustness of the findings to attrition and deviations from the experimental protocol. The primary finding of this reanalysis is that girls garnered substantial short- and long-term benefits from the interventions, particularly in the domain of total years of education. However, there were no significant long-term benefits for boys. These conclusions change little when allowance is made for attrition and possible violations of random assignment.

*I thank my advisors, Joshua Angrist, David Autor, and Jon Gruber, for their insightful comments and suggestions. I also thank Daron Acemoglu, Bruce Meyer, Michael Wolf, and the participants of the MIT Public/Labor Seminar and MIT Labor Lunch for helpful input. I am extremely grateful to Larry Schweinhart and Zongping Xiang of the High/Scope Educational Research Foundation, Frances Campbell, Elizabeth Pungello, and Richard Addy of the FPG Child Development Institute at University of North Carolina, Chapel Hill, and Craig Ramey of the Georgetown Center for Health and Education for their generous assistance in obtaining the Perry Preschool Program and Abecedarian Project data used in this study. This research used the Early Training Project, 1962-1979 [made accessible in 1988 as microfiche and numeric data files]. These data were collected by Susan Walton Gray, and are available through the Henry A. Murray Research Archive of the Institute for Quantitative Social Science at Harvard University, Cambridge, Massachusetts [Producer and Distributor]. Funding from the National Institute on Aging, through Grant Number T32-AG00186 to the National Bureau of Economic Research, is gratefully acknowledged. The author bears sole responsibility for the contents of this paper.

1 Introduction

The education literature contains dozens of papers showing inconsistent or negligible returns to publicly funded human capital investments (Hanushek, 1996). In contrast to these studies, several randomized preschool experiments report striking increases in short-term IQ scores and long-term outcomes for treated children (Schweinhart, et al., 2005; Campbell, et al. 2002; Gray, Ramsey, and Klaus 1982). These results have been highly influential and are frequently cited as proof of efficacy for many types of early interventions (for example, Cunha, et al., 2005). They have contributed to a widespread perception that the Head Start program - one of the centerpieces of American education policy - is effective and encouraged further research on preschool programs (Currie, 2001). The experiments underlie the growing movement for universal pre-kindergarten education (Kirp, 2005). They also play an important role in the debate over the optimal pattern of human capital investments, with all parties agreeing that early education is a crucial component of human capital policy (Krueger, 2003; Carneiro and Heckman, 2003).

This paper focuses on three prominent preschool evaluations: the Abecedarian Project, the Perry Preschool Program, and the Early Training Project. Beginning as early as 1962, these programs targeted disadvantaged African-Americans in North Carolina, Michigan, and Tennessee respectively. These projects stand out from others because they implement a random assignment research design - participants were randomly assigned to treatment (preschool) or control groups.[1] This randomization overcomes the problem of confounding that affects many observational studies.[2]

Following the initial group assignment, treated children in each experiment received several years of preschool education (intensity differed across programs). Intervention continued until treated children began regular schooling. After that point, further intervention was limited to data collection.[3] Children in both treatment and control groups received a series of standardized tests, beginning before age five and lasting through their teenage years. Researchers also conducted subject interviews and examined school and

[1] Two other preschool evaluations utilizing a random assignment research design exist. They are the Houston Parent-Child Development Center and the Milwaukee Project. Houston PCDC did not collect data on later life outcomes and experienced high rates of attrition. Milwaukee Project used extraordinarily small samples and suffered from a scandal involving one of its primary researchers.

[2] If parents are allowed to select whether their children receive an intervention, it is likely that the children receiving the intervention will differ in important ways from the children who do not receive it. In the context of preschool education, economists typically assume that children who attend preschool come from families that are more affluent or place a higher priority on education. Observational studies can therefore be misleading because factors other than preschool intervention may confound the results. In the context of Head Start, Currie and Thomas (1995) and Garces, Thomas, and Currie (2002) address the issue by including mother fixed effects when estimating the effect of the Head Start program on early and later life outcomes.

[3] One notable exception occurred. As discussed in Section (2), some treated Abecedarian children also received a schooling age treatment for several years.

government records to collect long-term follow-up data on academic, social, and economic outcomes.

Like all experiments, notable deviations from the intended protocol occurred in each study. In two experiments, attrition materialized before preschool treatment and during the collection of follow-up data. As a result, the initial randomization in treatment status was effectively contaminated. Logistical concerns in the Perry Preschool Program also prompted the reassignment of select children between treatment and control groups, further perturbing the randomization.

In addition to the departures from experimental protocol, serious statistical inference problems affect these studies. The experimental samples are very small, ranging from approximately 60 to 120. Statistical power is therefore limited, and the results of conventional tests based on asymptotic theory may be misleading. Furthermore, the large number of measured outcomes raises concerns about multiple inference: significant coefficients may emerge simply by chance, even if there are no treatment effects.[4] All of these issues - combined with a puzzling pattern of results in which early test score gains disappear within a few years and are followed a decade later by significant effects on adult outcomes - have created serious doubts about the validity of the experiments.

This paper has three related objectives. First, it implements a unified statistical framework to directly address concerns about sample size and multiple testing. Second, it simultaneously examines all three studies to estimate the long-term effects of preschool separately for both males and females. Finally, it performs a detailed analysis of potential threats to validity, including attrition, violation of random assignment, and clustering.[5]

The paper is organized as follows. Section (2) describes the data and specific details regarding each program's experimental design. Section (3) sets out the statistical framework and briefly discusses possible complications.[6] Section (4) presents results organized by outcome stage: pre-teen, teenage, and adult. Section (5) summarizes the main results and discusses possible explanations for the observed causal effects. Section (6) concludes. The results demonstrate that preschool intervention has significant effects on later life outcomes for females, including academic achievement, economic outcomes, criminal behavior, drug use, and marriage. The effect on total years of education is particularly strong. However, while treatment effects are sizable for females, they are minimal or nonexistent for males - a fact relevant to the design of optimal

[4]Both Currie and Thomas (1995) and Krueger (2003) cite these inference problems as primary reasons to doubt the validity of the experimental results.

[5]To my knowledge, I am the first independent researcher to analyze the micro data for all three programs.

[6]The complications are addressed in detail in Appendix A.

human capital policy. A thorough analysis of threats to validity, conducted in Appendix A, concludes that the main results are unaffected by reasonable assumptions regarding attrition, violation of random assignment, and clustering.

2 Experimental Background and Data Description

2.1 The Abecedarian Project

The Abecedarian Project recruited and treated four cohorts of children in the Chapel Hill, North Carolina area from 1972 to 1977. Children were randomly assigned to treated and control groups.[7] The treated children entered the program very early (mean age, 4.4 months). They attended a preschool center for eight hours per day, five days per week, 50 weeks per year until reaching schooling age. The program focused on developing cognitive, language, and social skills. Children in the control group received iron fortified formula, free diapers, and supportive social services when appropriate (Campbell and Ramey, 1994). Of the three preschool projects, Abecedarian was the most intensive.

The Abecedarian dataset contains 111 children. Researchers recruited 122 subjects, but 11 families declined or could not participate. Of the remaining 111 infants, 57 were assigned to the treatment group and 54 to the control group. Data collection began immediately and has continued - with gaps - through age 21.

Researchers gathered data from three primary sources: interviews with subjects and parents, program administered tests, and school records. Children received IQ tests on an annual basis from ages two through eight, and then once at age twelve and once at age fifteen.[8] Other standardized tests were also administered, but I focus on IQ scores for comparability across programs.[9] Researchers collected information on grade retention and special education at ages twelve and fifteen from school records. Data on high school graduation, college attendance, employment status, pregnancy, and criminal behavior come from an age 21 interview. Follow-up attrition rates are low for most outcomes, ranging from three to six percent in general.

[7] In fact, the experiment used a slightly more complex 4-way design. Children were assigned to one of four groups: preschool treated, preschool and schooling age treated, schooling age treated, and untreated. The schooling age treatment is potentially relevant: Currie and Thomas (2000) present evidence that higher quality primary schools enhance the long-term effects of Head Start. However, in this case the schooling age treatment - which included supplemental educational activities and biweekly home visits for three years - had a negligible effect, perhaps because it was not very intensive. It is therefore ignored for the purposes of this analysis. See Campbell and Ramey (1994) for further details.

[8] Instead of receiving IQ tests at ages six and seven, a single IQ test was administered at age 6.5.

[9] In the externally available Abecedarian dataset, test scores outside the 5th and 95th percentiles are truncated to the 2nd and 98th percentiles. Thus the mean IQ scores reported here differ slightly from the IQ scores in the previous Abecedarian literature.

2.2 The Perry Preschool Program

The Perry Preschool Program recruited and treated children in Ypsilanti, Michigan from 1962 to 1967. Children were randomly assigned to treated and control groups.[10] Treated children entered the program at age three and remained in it for two years.[11] The program implemented the ideas of Jean Piaget and focused on language skills, socialization, numbers, space, and time. Classes were based around activities, and teachers used conversations to help children reflect upon what they did. Children attended the program five mornings per week from October through May. Treated children also received one 90 minute home visit per week. Untreated children were interviewed for data collection, but received no other services.[12]

Perry researchers recruited 128 subjects in five waves. Following random assignment within each wave, pairs of children with similar IQ scores were swapped between treatment and control groups to equalize socioeconomic status and sex ratios across the two groups. A few children with working mothers were switched from the treatment group to the control group; this issue is addressed in Appendix A. Four children in the treatment group moved away before completing preschool, and one child in the control group died. Ultimately, the treatment group contained 58 children and the control group 65, for a total sample of 123.

Researchers gathered data from four primary sources: interviews with subjects and parents, program administered tests, school records, and criminal records. IQ tests were administered on an annual basis from entry until age ten, and once more at age fourteen. Information on special education, grade retention, and graduation status was collected from school records. Arrest records were obtained from the relevant authorities, supplemented with interview data on criminal behavior. Economic outcome data come primarily from interviews conducted at ages 19, 27, and 40. Follow-up attrition rates for most variables are generally low, ranging between zero to ten percent.

2.3 The Early Training Project

The Early Training Project occurred in Murfreesboro, Tennessee from 1962 to 1964. Two waves of three to four year old children were randomly assigned to treated and control groups. The treated children attended preschool for ten weeks during the summer, four hours per day. The program continued until the beginning

[10] The published Perry literature claims that children were matched in pairs based on initial IQ scores. One child from each pair was assigned to treatment, and the other to control. However, there is no record of the original pairing. I therefore conduct the analysis as if there were no matching of this type. If this assumption is violated, the estimated standard errors will be more conservative than necessary.

[11] One wave entered at age four and received treatment for only one year.

[12] This description is drawn mainly from Schweinhart, et al. (2005). Please see that reference for further details.

of school, for a total of two to three summers of preschool. Children received positive reinforcement in the classes and participated in activities focusing on motivation, persistence, and postponement of gratification. Treated children also received one 90 minute home visit per week for the duration of the program.[13] Control children received no treatment beyond interviews for data collection.

The Early Training Project initially gathered data on 92 children. Four children were disqualified for various reasons, leaving 88.[14] The Early Training Project differs from the other two experiments in its construction of the control group. Specifically, the study's control group consists of two distinct subsets: a local control group and a distal control group. Of the 88 children in the study, 61 lived in the town of Murfreesboro, and 27 lived in a different Tennessee town. The 61 children in Murfreesboro were assigned to the treatment group with approximately two-thirds probability and the local control group with approximately one-third probability. The 27 children in the distant town formed the distal control group.

The reliance on a distal control group was an unfortunate choice in the experimental design. The two towns were not initially comparable. For example, the distal town had a higher rate of AFDC enrollment. During the project's data collection phase, trends between the two towns diverged substantially. The local town's population grew almost 25 percent, while the distal town's fell several percent. Educational outcome data also suggest that the local and distal control groups are not interchangeable. Distal control females, for instance, display a significantly higher graduation rate than local control females. I therefore drop the distal control group from my analysis, and retain only those subjects who were truly randomly assigned. This choice results in a treatment group of 43, a control group of 18, and a total sample of 65. Since the treatment and control groups are unbalanced, statistical power is even weaker than the total sample size suggests.

Early Training Project data come from three primary sources: interviews with subjects and parents, program administered tests, and school records. IQ tests were given annually from ages four through eight, and again at ages ten and seventeen. Information on grade retention and high school enrollment comes from school records. Subject interviews provide data on post-high school education status and economic outcomes. No crime data were collected. Attrition rates for most variables are below ten percent; females in particular had virtually no attrition for many variables.

[13] Home visits continued for one year after the last summer school session.

[14] Pretreatment data on the disqualified children was retained.

2.4 Summary Statistics

Table 1 lists means and standard deviations of key variables for all three projects. The statistics highlight the degree to which these children are disadvantaged. Average IQs in the teenage years range from 93.2 to 77.6. In comparison, an IQ score of less than 70 is one criteria that the *Diagnostic and Statistical Manual of Mental Disorders - Fourth Edition* uses to define mild mental retardation. High school dropout rates range from 30 to 40 percent. In at least one sample, a majority of subjects have a criminal record. When drawing inferences regarding the results' external validity, it is important to note that the children studied are not representative of the average American child. Nevertheless, many of their attributes are not unusual for African-American youth in urban environments.[15]

3 Statistical Framework and Potential Complications

3.1 Statistical Framework

The random assignment process makes estimation of causal effects straightforward. The primary approach compares treated children (those that received preschool) to untreated children (those that did not) across a wide variety of outcomes. In general, this difference approximates both the effect of the treatment on the treated (ETT) and the intention to treat effect (ITT). The equivalence between ETT and ITT occurs in this case because virtually every child assigned to the preschool group attended preschool, and the programs were not open to children outside the preschool group. In the language of Angrist, Imbens, and Rubin (1996), almost every member of the sample was a "complier."[16]

To conduct inference, I compute Huber-White standard errors that are robust to heteroskedasticity. Although these standard errors are asymptotically consistent, the samples are quite small - some groups contain as few as ten individuals. The Huber-White standard errors may therefore be misleading, particularly since

[15]For example, Miller (1992) estimates that on any given day in 1991, 56 percent of African-American males aged 18-35 in Baltimore City were under some form of criminal justice supervision.

[16]It is conceivable that some children in the control group attended *different* preschool programs. However, this is unlikely. The families in these studies were relatively poor, so it would be difficult for most of them to afford private preschool programs. The predominant public preschool program, Head Start, did not begin until 1965, and it was initially a summer program. It therefore cannot have affected results for the Early Training Project, which ended in 1964, or the Perry Preschool Program, which had no summer session. In the latter case, the data show that fewer than 20 percent of Perry children attended Head Start, and these children were distributed fairly evenly between the treatment and control groups. The Abecedarian control children, however, may have received some Head Start schooling. It would be interesting to know whether any Abecedarian control children participated in Head Start, and how their outcomes differed from control children who did not. To my knowledge this information does not exist.

the underlying data is distributed non-normally in some cases.[17] To address this concern, I calculate *p*-values that do not rely on asymptotic theory or distributional assumptions.

Instead of a standard *t*-test, I implement a variant of the non-parametric permutation test (Yucesan, 1995). This procedure computes the null distribution of the test statistic and requires only three assumptions: random assignment, independence, and no treatment effect. For a given sample size N_k, I draw outcomes y_i^* from the empirical distribution of y_i without replacement.[18] I draw binary preschool assignments z_i^* with probability $p = 0.50$ (or $p = 0.67$ in the case of the Early Training Project) with replacement. For each sample, I calculate the *t*-statistic for the difference in means between treated and untreated groups. I repeat the procedure 10,000 times and compute the frequency with which the simulated *t*-statistics - which have expectation zero by design - exceed the observed *t*-statistic. If only a small fraction of the simulated *t*-statistics exceed the observed *t*-statistic, I reject the null hypothesis of no treatment effect.[19]

This test is similar to several well-known tests. If the preschool assignments z_i^* were sampled *without* replacement from the empirical distribution of z_i, this procedure would generally converge to Fisher's Exact Test for binary y_i.[20] Alternatively, if the outcomes y_i^* were drawn from the empirical distribution of y_i *with* replacement, the procedure would be analogous to bootstrapping under the assumption of no treatment effect (Simon, 1997). The procedure diverges from these two techniques because it attempts to reproduce the actual experiment as closely as possible. The procedure samples the outcomes y_i^* without replacement because the original sample is not a random sample of any larger population. It samples the preschool assignments z_i^* with replacement because the original assignments were drawn with replacement.[21]

The reported *p*-values are correct for tests conducted in isolation, but they do not address the issue of multiple inference. Because each study examines hundreds of outcomes, some outcomes should display significance even when no effect exists. Furthermore, the small samples ensure that significant results are necessarily of notable magnitude.

I address the issue of multiple inference in three steps. First, to minimize the degree of over-testing,

[17] Horowitz (2001) demonstrates that the performance of Huber-White standard errors can be very poor in small samples. Currie and Thomas (1995) and Krueger (2003) explicitly express concerns about the small Perry samples.

[18] Since these outcomes are drawn without replacement, and the sample size is N_k, in practice I simply use the original vector of y_i observations.

[19] Formally, I reject the hypothesis that the treatment has any distributional effect. For non-binary outcomes, it is theoretically possible that rejection occurs because treatment affects dispersion without affecting the mean. This seems unlikely. Furthermore, most of the outcomes of interest are binary, and anything that affects the variance of a Bernoulli random variable necessarily affects the mean as well.

[20] The procedure differs very slightly from Fisher's Exact Test in that Fisher's test rejects for small *p*-values while this test rejects for large *t*-statistics.

[21] Using alternative tests in which all sampling is done with or without replacement does not significantly affect the results.

I choose a specific set of primary outcomes based on a priori notions of importance. Next, I implement summary index tests in three broad areas: pre-teen, adolescent, and adult outcomes.[22] Finally, I control for multiple inference at the summary index level by computing Familywise Error Rate (FWE) adjusted p-values via the free step-down resampling method.

The set of primary outcomes includes grade retention, special education, high school graduation, college attendance, employment, earnings, government transfers, arrests, convictions or incarcerations, drug use, teen pregnancy, and marriage. This list appears long but represents only a small fraction of all available outcomes. Nevertheless, the total number of tested outcomes exceeds 40. I therefore implement summary index tests that pool multiple outcomes into a single test.

The summary index tests originate in the biostatistics literature (see O'Brien, 1984). They are robust to over-testing because the probability of a Type I error does not increase as additional outcomes are added to a summary index. They are also potentially more powerful than individual level tests - multiple outcomes that approach marginal significance may aggregate into a single index that attains statistical significance.[23]

To implement these tests, I demean all outcomes and convert them to effect sizes by dividing each outcome by its standard deviation. This conversion normalizes outcomes to be on a comparable scale. I also switch signs where necessary so that the positive direction always denotes a "better" outcome. I then create a new variable, $\overline{s}_{ij}$, that is the mean of the normalized, demeaned outcomes.[24] Thus $\overline{s}_{ij} = \sum_{k \in \mathbb{K}_{ij}} w_k \frac{y_{ijk} - \overline{y}_{kj}}{\sigma^y_{jk}}$, where k indexes outcomes within area j, K_{ij} is the total number of non-missing outcomes for observation i in area j, $\mathbb{K}_{ij}$ is the set of non-missing outcomes for observation i in area j, and w_k is the outcome weight (weights are normalized to sum to one). I then regress the new variable, $\overline{s}_{ij}$, on treatment status to estimate the effect of preschool on area j. Any missing outcomes are ignored when creating $\overline{s}_{ij}$. This procedure therefore uses all the available data, but it weights outcomes with fewer missing values more heavily.[25]

[22] Grouping instead by type of outcome - e.g. academic, social, economic - does not substantially alter the conclusions.

[23] Conceptually, it is useful to consider an underlying latent index - e.g., human capital at a given age - that is expressed through multiple measures (e.g., years of education, employment, earnings, criminal record). We want to know whether preschool affects the latent index, but there are two sources of random error to contend with. First, there is error that arises from the random assignment procedure - the latent index will not be perfectly balanced across treatment and control groups in any finite sample. Second, there is random error in each outcome measure - individuals with the same latent index value may realize different values for any given outcome. Summary index tests can reduce the second source of error by combining data from multiple outcome measures.

[24] The variable $\overline{s}_{ij}$ is also weighted by the inverse of the outcome covariance matrix; the weight on each outcome is equal to the sum of its row entries in the inverted covariance matrix. This weighting increases efficiency by ensuring that outcomes which are highly correlated with other outcomes receive less weight (or, alternatively, outcomes that represent new information receive more weight).

[25] An extreme case illustrates this point. Consider an example in which one outcome is missing data for every single observation. In that case, the outcome never enters into $\overline{s}_{ij}$ for any observation, and does not affect the estimation results. An alternative

Each summary index consolidates several individual tests into a single index. However, there are still nine summary tests per gender. I therefore calculate FWE adjusted p-values for all summary index tests. Suppose that *K* hypotheses, $H_1, H_2, ..., H_K$, are tested. The Familywise Error Rate (FWE) is the probability that at least one of the *K* hypotheses in the family is rejected.[26] For summary index tests, the family of tested hypotheses is the set of nine summary index tests performed for each gender.

To adjust for FWE, I implement the free step-down resampling method (Westfall and Young, 1993). This algorithm is more powerful than simpler FWE adjustments, such as the Bonferroni Correction, because it incorporates dependence between outcomes and sequentially removes hypotheses from the family being tested as they are rejected. An example may aid the interpretation of the adjusted p-values. Consider the smallest unadjusted general effect p-value, which occurs for teenage Perry females (Table 4). The unadjusted p-value is approximately 0.000. The corresponding adjusted p-value, calculated via the free step-down resampling method for the entire family of female summary tests, is $p = 0.003$. Suppose we simulate the female data 10,000 times under the null hypothesis of no treatment effect. If we compute an entire set of summary effect p-values for each simulation, then the *minimum* p-value of that set will be less than or equal to the unadjusted p-value of 0.000 approximately 0.3 percent of the time. For unadjusted p-values that are above the family's minimum p-value, the family of tests effectively decreases. A monotonicity enforcement performed at the end of the procedure ensures that larger unadjusted p-values always correspond to larger adjusted p-values. The code for this procedure is detailed in Appendix B.

3.2 Complications

Several complications, analyzed in-depth in Appendix A, threaten the validity of the results. A quick summary of the complications and their resolutions follows.

Attrition is present in all three preschool experiments. If this attrition is caused by treatment status, systematic differences unrelated to the treatment could emerge between the two groups. In these experiments,

estimator, detailed in Kling and Liebman (2004), simultaneously estimates the coefficients for all outcomes in a given area using a seemingly unrelated regressions (SUR) model. The general effect is computed as the mean of the coefficients in that area, and the estimate's variance is calculated using the coefficient variance-covariance matrix from the SUR model. However, this estimator drops an observation if it is missing for any outcome (with no missing outcomes, the two procedures are equivalent). Since neither test is superior on a priori grounds, I experiment with both. They return similar results, except for a few cases in which the SUR estimator is affected by a large number of missing observations. I therefore report results for the summary index estimator.

[26] Note that the FWE adjustment is not the same as a joint test of the hypothesis of no effect for any outcome. If a joint test rejects, we can only conclude that at least one null hypothesis is false. If the adjusted p-value rejects, we can conclude that the specific null hypothesis being tested is false. A joint test is generally more powerful, but, when it rejects, the adjusted individual test yields more information.

the direction of the induced bias is ambiguous. To address the attrition problem, I impute values for key outcomes among missing individuals and examine "worst case" scenarios. Under reasonable assumptions, these imputations do not qualitatively change the paper's central conclusions.

Another complication is violation of the original random assignment. The most serious case occurred in the Perry Preschool Program; for logistical reasons, several children with working mothers in the treatment group were switched to the control group. Perry researchers did not record the identities of these children. If children with working mothers perform differently than the average child, these swaps could induce bias. I address this issue by conditioning outcomes on initial maternal employment status. I also study an entire range of possible switches that could have occurred and examine the sensitivity of the estimates to these switches. Again, the main results are unchanged.

A final complication is the possibility of dependence between observations, or clustering. In these experiments, the possibility of classroom peer effects and the systematic assignment of siblings to identical treatment groups are reasons for concern. If the peer effects or intra-family correlations are strong, the standard errors could be too small. I address the problem by estimating the results on a dataset of class-by-year means and by dropping siblings from the sample. The clustering adjustments do not substantially affect key results.

4 Results

4.1 Pre-Teen Outcomes

Preschool significantly raises early IQ scores in all experiments. It also consistently reduces early grade retention and special education placement for females, but has limited effects on grade retention and special education for males.

Table 2 reports effects on pre-teen IQ scores. Like all tables in this section, it presents results for both genders. For each gender, the first column reports coefficients and standard errors, the second column reports control group means, the third column reports non-parametric p-values, and the fourth column reports sample size. The last column in each table tests for differences between female and male treatment effects.

All projects demonstrate similar effects on test scores at early ages. In each project, there is a large and significant IQ effect for at least one gender upon completion of preschool. Females continue to display a significant IQ effect at age ten in both the Abecedarian and Early Training Projects. Males, however,

experience no significant IQ effect in any project at age ten.

The similarity in early IQ effects across programs occurs despite their differing intensity levels. By age five, the Abecedarian, Perry Preschool, and Early Training programs exposed children to approximately 8,000, 850, and 700 hours of preschool education respectively.[27] Nevertheless, a treatment effect that peaks at roughly ten to fifteen IQ points emerges in all three programs during the preschool years.

The results in Table 3 suggest that the early IQ gains translate into better performance in primary school.[28] Female grade retention falls by 20 to 30 percentage points in all three programs, with *p*-values ranging from 0.08 to 0.16. Female special education placement falls significantly in the Perry program (26 percentage points, $p = 0.06$) but not in the Abecedarian or Early Training programs. Males in the Abecedarian program experience a 19 percentage point decline in grade retention ($p = 0.14$) and a 27 percentage point decline in special education placement ($p = 0.06$). However, males in the Perry and Early Training programs demonstrate *increases* in grade retention of approximately 8 to 10 percentage points and no notable decrease in special education placement.

Table 4 reports summary index results by outcome stage and experiment. At the pre-teen stage, preschool significantly improves outcomes for females in the Abecedarian and Perry programs, with summary effect size increases of 0.45 and 0.54 respectively. Early Training females experience a summary effect size increase of 0.38; the coefficient approaches significance. Males, in contrast, do not experience consistent gains in pre-teen outcomes. Abecedarian males realize a significant summary effect size increase of 0.42. However, Perry and Early Training males experience summary effect size increases of 0.15 and 0.14 respectively; neither result approaches significance.

Gender differences in treatment effects emerge by age ten. The female IQ effects at age ten are significantly higher than the male IQ effects in both the Perry and Early Training programs. Females also experience greater drops in grade retention than males in both the Perry and Early Training programs, and the differences approach significance. Most importantly, for every experiment the summary female pre-teen effect is higher than the summary male pre-teen effect; the difference approaches marginal significance in the Perry Preschool Project.

[27] Currie and Neidell (2004) present evidence that higher spending increases the effects of Head Start. Although initial differences are minimal, the Early Training Project does have the lowest number of significant long-term outcomes. However, this is partially due to the relatively small samples in the Early Training Project.

[28] For Perry Preschool, the grade retention variable may contain some information on teenage grade retention. For the Early Training Project, both the grade retention and special help variables may contain some information from teenage years. For these variables, it was not possible to isolate pre-9th grade outcomes in the data.

Although preschool positively affects pre-teen outcomes, the implications for long-term success are unclear. A short-term IQ gain may not result in any long-term economic benefit, and decreased grade retention at an early age may not affect graduation rates a decade later. For example, Currie and Thomas (1995) and Garces, Thomas, and Currie (2002) conclude that, for African-Americans, Head Start initially boosts test scores but does not have any lasting effect on academic achievement or economic outcomes. Conversely, diminishing effects on standardized tests may mask improvements in crucial non-cognitive skills that affect earnings and achievement (Heckman and Rubinstein, 2001). The next subsections focus on long-term teenage and adult outcomes.

4.2 Teenage Outcomes

In the teenage years, early intervention significantly improves high school graduation, employment, and juvenile arrest rates for females. However, it has no significant effect on male outcomes.

Table 6 presents program effects on teenage academic outcomes, including IQ scores and high school graduation rates. By age 14, initial IQ effects dissipate in all three programs. Only one IQ coefficient is statistically significant - Abecedarian males at age 15 ($p = 0.09$) - and in no case does the estimated coefficient exceed five IQ points. However, the negligible IQ effects belie strong gains among females for several important teenage outcomes.

High school graduation effects for females are sizable. Females display increases in high school graduation rates (or decreases in drop out rates) of 23 percentage points in the Abecedarian Project, 49 percentage points in the Perry Preschool Program, and 29 percentage points in the Early Training Project. The Perry result is highly significant ($p < 0.001$). The Abecedarian and Early Training results achieve or approach marginal significance ($p = 0.09$ and $p = 0.11$ respectively).[29]

In contrast, the high school graduation effects for males are weak or negative. Graduation rates *decline* by 10 and 6 percentage points for Abecedarian and Perry males respectively. Early Training males are 10 percentage points less likely to drop out, but the effect is not statistically significant.

Table 7 presents results for teenage economic and social outcomes. Females display positive economic effects from preschool as teenagers. In Perry Preschool, treated females have teen unemployment rates that are 31 percentage points lower than untreated females ($p = 0.03$). Treated females also receive approxi-

[29]The relative insignificance of the Early Training results is primarily a result of the relatively small sample size. The estimated coefficient is larger than the Abecedarian coefficient, but with only ten females in the Early Training control group it is difficult to conduct accurate statistical inference.

mately 1,600 dollars less in annual government transfers at 19 ($p = 0.04$). Early Training females are 13 percentage points more likely to have worked as teens, although the effect is not significant. Males, in comparison, derive no significant economic benefits from preschool during their teenage years. Unemployment among Perry male teens is only 2 percentage points lower. Treated male teens in the Early Training Project are 6 percentage points *less* likely to have ever worked.

The preschool programs have moderate effects on teen motherhood. Abecedarian females report teen pregnancy rates that are 21 percentage points lower; the effect approaches marginal significance ($p = 0.13$). Teen pregnancy rates for Perry females are 19 percentage points lower, but the effect is insignificant. Neither Abecedarian nor Perry males experience a significant decline in the probability of teen parenthood.

Early intervention has a significant effect on female teen criminal behavior. It reduces the probability of a juvenile record by 34 percentage points for Perry females. However, this significant result ($p = 0.01$) is not mirrored among males. Perry males demonstrate an insignificant 8 percentage point reduction in the probability of arrest before age 20.

Overall, preschool has a consistent, positive effect on female teen outcomes. Teenage summary effects increase by 0.42, 0.61, and 0.55 respectively for females in the Abecedarian, Perry, and Early Training programs (see Table 4). The summary effect is highly significant for Perry females ($p < 0.001$) and retains significance when adjusted for multiple testing. The summary effect is also significant for Abecedarian females ($p < 0.05$), but only marginally significant for Early Training females. However, preschool has no significant effect on male teen outcomes. Summary effects increase for males by only 0.16, 0.04, and 0.10 respectively in the Abecedarian, Perry, and Early Training programs. No male summary effect approaches statistical significance.

During the teenage years, it is clear that females benefit more than males from early intervention. The female-male difference in high school graduation effects is significant in the Abecedarian Project ($t = 1.80$) and the Perry Preschool Program ($t = 3.32$). Large female-male differences also emerge among Perry teens for effects on unemployment ($t = -1.60$), criminal behavior ($t = -1.54$), and government transfers ($t = -1.96$). At the summary index level, Perry females benefit significantly more than Perry males ($t = 3.32$). For the other two experiments, female summary effects are at least 0.25 standard deviations higher than male summary effects, although the differences are not significant. With the exception of Abecedarian IQ test scores, every reported teen effect is more positive for females than for males.

4.3 Adult Outcomes

At the adult stage, preschool significantly raises college attendance rates for females and appears to improve female economic outcomes and reduce criminal behavior. The effects for males, however, are weak and inconsistent. There is evidence of a modest positive effect on male economic outcomes, but it is accompanied by evidence of a negative effect on male college attendance and a mixed effect on male criminal behavior.

Table 8 reports treatment effects on college attendance. Preschool appears to increase the probability of college attendance for females. Abecedarian females report college attendance rates 29 percentage points higher than their control counterparts. This result is statistically significant ($p = 0.02$). Perry female college attendance rates increase by 16 percentage points, and Early Training females are 12 percentage points more likely to obtain post-high school education, although neither effect is significant.[30]

However, preschool does not appear to increase college attendance for males. Abecedarian males display a 15 percentage point increase in college attendance rates, but the effect is insignificant. Perry males are 1 percentage point less likely to attend college, and Early Training males report dramatically lower rates of post-high school education (49 percentage points lower).[31] The negative effect for Early Training males is highly significant ($p = 0.005$).[32]

Table 9 reports results for adult economic outcomes. Preschool has a weak but positive effect on female economic outcomes. Abecedarian women are 10 percentage points more likely to be employed at age 21. Perry females are 26 percentage points more likely to be employed at age 27 ($p = 0.08$). However, this effect disappears by age 40. Perry females earn more at ages 27 and 40 than their control counterparts (approximately 2,600 and 3,500 dollars respectively), but the effects are insignificant.[33] Early Training

[30]Post-high school education is defined as college, vocational school, or employer sponsored education/training. For either gender, limiting the outcome to just college attendance produces coefficients of similar magnitude and significance.

[31]In cases where there is overlap, my results are similar - but not exactly identical - to the results reported in Gray, et al. (1982). The discrepancy arises because the dataset that Dr. Gray provided to the Murray Research Center does not exactly match the description of the dataset used in Gray, et al. (1982). Dr. Gray passed away several years ago, so it is unlikely that we can ever fully resolve these minor discrepancies.

[32]The most likely reason for this negative finding is multiple testing. Two other possibilities are attrition bias and negative peer effects. A detailed examination reveals both of these explanations to be unlikely. Further discussion is available from the author upon request.

[33]For both females and males, the coefficient on monthly earnings at 27 has a much higher *t*-statistic than the coefficient on annual earnings at 27. This difference arises because the monthly earnings coefficient is between $\frac{1}{4}$ to $\frac{1}{6}$ the magnitude of the annual earnings coefficient, rather than the expected $\frac{1}{12}$. There is no a priori reason to believe that one measure is clearly superior to the other. However, the annual earnings measure does have a lower standard deviation than the annualized monthly earnings measure. More importantly, using annual earnings at 27 instead of monthly earnings at 27 produces an estimate that is consistent with the estimated earnings differentials at age 40 using either monthly or annual measures. The implied earnings effect using annual reported earnings at age 27 is 19 percent of the control mean, while the implied earnings effect using monthly reported earnings at age 27 is 59 percent of the control mean. The implied earnings effects using annual and monthly reported earnings at age 40 are 24 and 17 percent of the control means respectively. The reported monthly earnings at age 27 therefore appear

females are less likely to receive welfare at age 21, but are also less likely to receive income from work at the same age (neither effect is significant). It is possible that for Abecedarian and Early Training women, potential employment effects at age 21 are masked by increased college attendance rates. In that sense, employment data at a later age would be preferable. However, controlling for college attendance when estimating the employment effect does not appreciably change the coefficients for either program.

For males, there is no consistent evidence that preschool interventions improve long-term economic outcomes. Abecedarian males achieve an employment rate 19 percentage points higher than their untreated counterparts, but Perry males see virtually no effect on employment at age 27. Perry males do report increases in annual earnings of approximately 2,400 and 6,200 dollars at ages 27 and 40 respectively. However, all of these effects are insignificant. Perry males at age 40 experience a positive employment effect of 20 percentage points. This effect approaches statistical significance ($p = 0.11$). Early Training males, however, are *less* likely to receive income from work at age 21.

Table 10 presents effects on adult social behavior. Treated females report improvements for several measures of criminal behavior. Abecedarian females are 32 percentage points less likely to use marijuana ($p < 0.01$). However, Abecedarian does not significantly reduce conviction or incarceration rates for females by age 21.[34] Perry females have 86 percent fewer lifetime arrests (a reduction of 1.95 arrests, $p = 0.01$), though they are only 15 percentage points less likely to have a criminal record.

Treated males, in contrast, do not show significant improvements for any reported indicator of criminal behavior. Abecedarian males are slightly less likely to be convicted by age 21 or to use marijuana. Perry males are 2 percentage points less likely to have a criminal record at age 27. Perry males have 38 percent fewer lifetime arrests at age 27, but the effect only approaches marginal significance (a reduction of 2.31 arrests per capita, $p = 0.13$). The "hard" drug usage rate is 20 percentage points *higher* for Perry males, an effect which attains statistical significance ($p = 0.07$).[35]

There is some evidence that preschool affects marriage rates.[36] At age 27, Perry females have a significantly higher marriage rate than untreated females. The 32 percentage point increase represents a 382

anomalous. Nevertheless, for completeness I take an average of the two measures when calculating the summary index estimator.

[34] It is tempting to assume that Abecedarian females experience no significant reduction in non-drug related criminal behavior because their underlying arrest rate is much lower than Perry females. However, this assumption is incorrect, because the Abecedarian data measures *convictions* while the Perry data measures *arrests*. Clarke and Campbell (1998) report that 43 percent of the Abecedarian sample have criminal records at age 21. 51 percent of the Perry sample have an arrest record at age 19, so the two numbers are quite comparable, particularly since the Perry sample has a higher proportion of males. Clarke and Campbell find no effect of early intervention on criminal records.

[35] This detrimental effect has the highest significance level of *any* of any major later life outcome measures for Perry males.

[36] Perry is the only program to date that surveys participants late enough to collect meaningful marital statistics.

percent rise over the control group's base rate ($p < 0.01$).[37] Perry males, however, have the same marriage rate at 27 as their control counterparts.[38]

Overall, females benefit from early intervention as adults. In the Abecedarian and Perry Preschool programs, females display positive general effects of 0.45 and 0.36 standard deviations respectively (see Table 4).[39] Both results are statistically significant ($p < 0.01$ and $p = 0.02$ respectively), and the Abecedarian effect is robust to FWE adjustments. However, Early Training females demonstrate no general treatment effect as adults. This could be a result of the Early Training Project's relatively short intervention program, or it could be due to low statistical power.

Unlike females, males demonstrate little evidence of positive treatment effects as adults. Summary effects for Abecedarian and Perry males increase by 0.31 and -0.02 standard deviations respectively. The Abecedarian result approaches significance, but the Perry result does not. Early Training males experience a *decline* of 0.65 standard deviations in the adult summary index. This significant decrease ($p < 0.05$) is primarily driven by low college attendance rates.

Several female treatment effects are significantly higher than corresponding male effects, although the effect heterogeneity is less pronounced than during the teenage years.[40] The female-male treatment effect difference is significant for drug use and marriage among Perry participants ($t = -2.07$ and $t = 2.00$) and post-high school education among Early Training participants ($t = 2.35$). The difference in female-male summary effects is also significant in the Early Training Project. For drug use and post-high school education, the significance is primarily the result of negative male treatment effects rather than positive female treatment effects. Nevertheless, it still constitutes evidence of greater benefits for females.[41]

[37] Interestingly, Schweinhart, et al. (2005) show that by age 40 the treated females' marriage rate is only 6 percentage points higher than the control females' rate. Part of the increase is due to divorces in the treatment group, and part is due to marriages in the control group.

[38] Again, Schweinhart, et al. (2005) show an interesting twist for males at age 40. Treated males are more likely to be married at age 40, but the entire increase is due to a larger fraction of treated males who have divorced and married multiple times. The fraction of treated males who have only married once is actually slightly lower than the fraction of controls who have only married once. It is unclear whether this pattern should be counted as a "positive" or "negative" effect.

[39] The Perry Preschool summary index also includes monthly income.

[40] The effect heterogeneity is reduced primarily because of a decline in the general effect size for females.

[41] The female coefficients are centered around a higher mean, so even in the face of adverse shocks they do not become negative and significant. The male coefficients, in contrast, are centered around a lower mean, and are more likely to display negative, significant effects simply due to chance.

5 Discussion

A clear pattern emerges from a detailed examination of preschool treatment effects by gender: females display significant long-term effects from early intervention, while males show weaker and inconsistent effects.[42] Treated females show particularly sharp increases in high school graduation and college attendance rates, but they also demonstrate significant positive effects for economic outcomes, criminal behavior, drug use, and marriage.

In contrast to females, males do not appear to derive lasting benefits from early intervention. No positive, long-term outcome achieves statistical significance for males, although one, employment at age 40 for Perry males, comes close. This aggregate performance is disappointing when considering the number of outcomes tested; even with a minimal treatment effect, positive and significant results are likely to occur several times just by chance. In fact, the only significant, long-term results for key male indicators are negative.

Figure 1 presents a visual summary of the female-male treatment effect heterogeneity for long-term outcomes. This figure plots *t*-statistics for all of the reported teenage and adult coefficients across all experiments. Each point corresponds to the *t*-statistic for a single outcome, and all outcomes have been recoded so that the positive direction always corresponds to a "better" outcome. The first column of points plots male *t*-statistics, and the second column plots female *t*-statistics. It is clear upon visual inspection that the distribution of female *t*-statistics is centered well above the distribution of male *t*-statistics.

The third column of points plots a set of male *t*-statistics generated by randomly assigning treatment status to males. This procedure guarantees that any significant "treatment effects" visible in the column are simply due to chance. The procedure is equivalent to sampling random draws from the *t*-distribution, except that it preserves the inherent correlation structure between *t*-statistics within each experiment.[43] To construct this column, I randomly generate a set of treatment assignments and compute the corresponding *t*-statistics. I then plot these *t*-statistics.

A comparison of the first and third columns demonstrates that the distribution of male *t*-statistics is difficult to distinguish from a draw of randomly generated *t*-statistics. The maximum value in the third column exceeds the maximum value in the first column, but the first column has more *t*-statistics clustered

[42] Several researchers, most recently Heckman (2005), have noted the possibility of heterogeneous treatment effects by gender in the context of Perry Preschool. However, there has been no statistical analysis of this difference, nor would it be possible to draw any strong conclusions regarding treatment effect heterogeneity by gender from Perry Preschool alone.

[43] For example, if the Abecedarian high school graduation *t*-statistic is large, then it is likely that the Abecedarian college attendance *t*-statistic will also be large. Therefore, patterns of large or small *t*-statistics are more likely to occur in this data than would be expected in a set of 29 independently sampled *t*-statistics.

above 1.5. In either column, a case can be made for positive treatment effects by focusing on the subset of outcomes near the top. This fact highlights the importance of correcting for multiple testing.

A formal analysis examines summary index FWE p-values and aggregates all long-term outcomes into a single summary index. In comparison to females, each of the nine male summary index coefficients is lower, often by a large margin. Female general effects attain significance for pre-teen, teenage, and adult outcomes in both the Abecedarian and Perry Preschool programs.[44] With the exception of Abecedarian teens, all of these effects at least approach marginal significance after FWE adjustment. Male general effects attain significance only for Abecedarian pre-teens and Early Training Project adults (the latter effect is negative). However, after adjusting for multiple testing, only the Early Training Project adult general effect achieves marginal significance.

A summary test that pools all teen outcomes together across experiments finds an overall effect size of 0.53 for females (standard error of 0.14) and 0.08 for males (standard error of 0.19). The gender difference is statistically significant at the 5 percent level. A summary test that pools all adult outcomes together across experiments finds an overall effect size of 0.27 for females (standard error of 0.09) and -0.05 for males (standard error of 0.11). The gender difference is again statistically significant at the 5 percent level. Of course, we can never reject an arbitrarily small effect for males, and precision is limited by the relatively small samples. Perhaps real male effects exist but are masked by the standard errors. Nevertheless, the results indicate that any positive male treatment effect is modest at best.

The reported heterogeneity in treatment effects by gender is consistent with several previous findings in the non-experimental literature. For example, Oden, et al. (2000) report that Head Start participation significantly raises high school graduation rates and lowers arrest rates for females. However, no significant effect is found for males. The results also parallel findings in other areas of the human capital literature. Kling and Liebman (2004) report that the Moving to Opportunity program improves educational outcomes and mental health for females, but appears to have *negative* effects on male participants. Abadie, Angrist, and Imbens (2002) find that the Job Training Partnership Act (JTPA) significantly increases female earnings at all quantiles, including a 35 percent increase at the lowest quantile. However, the JTPA has no significant effect on males at any quantile below the median, and the proportional effect never exceeds 12 percent.

A variety of explanations can account for the observed gender differentials. Testing these explanations

[44] Pre-teen and teenage female general effects are of notable magnitude in the Early Training Project, but do not attain significance because of the limited sample size.

is beyond the scope of this paper and its data. Nevertheless, a quick summary of possibilities is in order.

One likely possibility is that child development differs between boys and girls. Many researchers believe that girls develop faster than boys. For example, a recent longitudinal study of Australian children found that preschool age females outperform their male counterparts in the physical, social/emotional, and learning domains (Australian Institute of Family Studies, 2005). Evidence is also mounting that education has a greater impact at later stages of development. Fredriksson and Öckert (2005) discover that Swedish children who start school later get more education than their younger peers. This effect is more pronounced for children from weaker socio-economic backgrounds. If additional maturity enhances the effect of schooling, and girls mature faster than boys, then girls should benefit more than boys from early intervention.[45]

Disadvantaged females may also experience different obstacles than disadvantaged males, and non-cognitive skills developed in preschool might address the obstacles that females face more effectively. A possible example is the role of teen pregnancy in high school dropouts. Males cannot get pregnant, so any effect of preschool on teen pregnancy only benefits females. If teen pregnancy increases the likelihood of dropping out, preschool will have a greater effect on female educational attainment than male educational attainment. However, the data invalidate this particular explanation. Even if pregnancy caused a one-for-one increase in high school dropout status, the observed pregnancy effect still could not explain a majority of the female high school graduation effect. Nevertheless, other differences in obstacles faced by males and females may play important roles.

A third possibility is the existence of a selection effect. "Female" families participating in the program may differ from male families along unobserved dimensions. Gender is typically thought of as randomly assigned, but families with girls may be more or less likely to enroll in preschool programs (the Perry sample, for example, includes significantly more males than females). However, this fact need not invalidate the external validity of the results. If the same selection factors operate in the general population, then the reported female-male differences will be applicable to many preschool programs with voluntary participation.

Finally, recent research has established that students may perform better when taught by teachers of the same gender. For example, Dee (2005) presents evidence that middle school children are perceived as less disruptive and more attentive when the teacher is of the same gender. To my knowledge, all of the preschool

[45]Note that this hypothesis need not be inconsistent with the hypothesis that early intervention is more effective than later intervention. Since free public schooling past age 5 is universally available, later interventions are implicitly being performed on the intensive margin. Early interventions, in contrast, are often performed on the extensive margin. It is therefore possible that early interventions might be more effective, even if education is more effective for older children.

teachers in each experiment were female. If preschool age children also perform better when taught by adults of the same sex, then we might expect females to benefit more from early intervention than males.

6 Conclusion

This paper conducts a robust reanalysis of the influential experimental preschool literature. It partially confirms previous findings, presenting strong evidence that females benefit from early intervention. Significant female effects appear in the domains of criminal behavior, marriage, and economic success, but the most consistent improvement is an increase in total years of schooling. These results are robust to reasonable concerns regarding attrition, violation of random assignment, and clustering. Many female results also remain significant after adjusting for multiple inference.

For males, however, there is no evidence of positive, long-term preschool treatment effects. Most coefficients are insignificant, and several of the significant coefficients imply an adverse effect. The overall pattern of male coefficients is consistent with the hypothesis of a minimal treatment effect at best - significant effects go in both directions and appear at a frequency one would expect simply due to chance.[46]

The observed differences between female and male treatment effects are significant in a number of cases, particularly with respect to total years of education. The difference in overall long-term female and male summary indices is also significant. However, additional research with new data is necessary to determine the exact magnitude of the female-male treatment effect differential, and to discover whether males derive modest benefits from preschool intervention or no benefits at all.

In the context of the current human capital literature, this paper makes clear several points. Foremost, intensive preschool intervention does positively affect later life outcomes, at least for disadvantaged African-American females. However, there is no evidence of strong long-term preschool benefits for males. This fact suggests that investments in early education alone may not dramatically improve opportunities for disadvantaged males. The indicated treatment effect heterogeneity also calls into question the external applicability of experimental estimates. If treatment effects vary by gender, it is plausible that they may also vary by race or class. Richer variation in sample demographics is necessary for the design of optimal human capital policy. As Hanushek (2003) suggests, financing broader experimental research on human capital investments may well yield the highest return today of any human capital policy.

[46]Previous research has overlooked this result because there has been no systematic analysis by gender across experiments, nor has anyone applied a unified statistical framework that is robust to problems of multiple inference.

Appendix

A Assessing Threats to Validity: Attrition, Violation of Random Assignment, and Clustering

This paper reports significant long-term effects for females in the domains of educational achievement, criminal behavior, marriage, and economic success. The experimental design alleviates concerns about confounding variables, but there remain several issues specific to the individual studies that could cause the treatment and control groups to be systematically different in ways unrelated to the treatment, or cause statistical tests to over-reject. A careful examination of these issues is necessary before long-term effects for females, and the larger body of experimental preschool research in general, can be readily accepted.

The first problem is attrition, which occurs in the Abecedarian Project and the Perry Preschool Program.[47] The second problem is the intentional exchange of children between groups. This issue occurs only in the Perry Preschool Program. The final issue is clustering, or correlation between individual observations, which occurs primarily in the Perry Preschool Program. I find that the key results remain unchanged after accounting for these problems.

A.1 Attrition

Random attrition reduces statistical power but does not cause bias. Non-random attrition is acceptable if it is unrelated to treatment status - it will not induce systematic differences between the treatment and control groups, and estimated effects remain internally valid.[48] Therefore, our only concern is attrition that is caused by assignment status.

Attrition of two types occurs in the preschool experiments. The first type, which I refer to as follow-up attrition, occurs when individuals initially in the sample cannot be located for follow-up interviews, testing, or records collection. The second type, which I refer to as pre-treatment attrition, arises when individuals drop out after receiving their assignment but before entering the sample. In practice, the first type often

[47] Only one female is missing for most Early Training Project results. There is no documented evidence of attrition occurring after the random assignment but before data collection. Because attrition is almost non-existent in this study, and because the study found few significant effects to begin with, no attrition analysis is performed for the Early Training Project.

[48] Non-random attrition of this type can still affect the external validity of estimated treatment effects, but this caveat applies to all studies whose participants are not randomly drawn from the relevant population.

receives more attention than the second, perhaps because the missing data is readily apparent. Nevertheless, the two types are fundamentally similar.

If attrition is present, the direction of bias it produces is ambiguous. Most of the pre-treatment attrition occurs among children assigned to the treatment groups. In this case, we might expect a positive bias if families that care least about education are the ones refusing treatment.[49] Follow-up attrition affects both treated and control children. The leading causes of follow-up attrition are death and inability to locate the subject. Signing this bias with certainty is infeasible. However, it is notable that more control children died than treated children. If control children who die are especially poor or disadvantaged, attrition from death would attenuate a positive treatment effect. If successful subjects are likely to move out of state, then attrition from movement would also attenuate a positive treatment effect. We therefore might expect follow-up attrition to exert a negative bias. We cannot accurately guess the direction of the overall bias.

A.1.1 Abecedarian

The Abecedarian Project lost eleven children to pre-treatment attrition. Seven treatment group families and one control group family withdrew upon receiving their group assignments. Two control group children received preschool treatment due to medical conditions requiring close supervision; these children are not present in the dataset. An additional seven children were lost to follow-up attrition for most outcomes. One treatment male, one treatment female, and two control females died early in life. Three additional subjects are not present for various reasons: one treatment female had a seizure disorder, one control female withdrew for family related matters, and one treatment female declined to participate in the age 21 interview.[50]

Table 11 reports estimates for key outcomes under a variety of attrition assumptions. The analysis focuses on females, because males suffer less attrition and demonstrate no significant effects.[51] Columns (1) and (2) focus on follow-up attrition only. Of the six missing females, four dropped out for medical reasons unlikely to be affected by treatment status (three deaths and one seizure disorder).[52] The analysis

[49] On the other hand, it is possible that some treatment families pulled their children out because they felt they could offer a better experience at home. Depending on the characteristics of these families, this explanation could lead to a negative bias.

[50] The information regarding attrition comes from Campbell, et al. (2002), Clarke and Campbell (1998), Campbell and Ramey (1994), and Ramey, Yeates, and Short (1984).

[51] Under extreme assumptions regarding missing values, some results for males could attain marginal significance. Such results would not constitute compelling evidence of a male treatment effect.

[52] All of these deaths occurred at an early age - generally less than one year. One might hypothesize that preschool affects infant death rates, particularly those resulting from accidents or infectious diseases. CDC data indicates that the magnitude of this effect would be trivial. Of the top causes of black postneonatal death in 1979 - which account for almost 70 percent of total black postneonatal deaths - preschool attendance could only affect accidents, homicide, pneumonia, bronchitis, viral infections, and meningitis to a significant degree. These causes account for only 19 percent of postneonatal deaths (Hoyert, Kochanek, and

therefore explores imputations for the two females that specifically chose not to participate in follow-up surveys.[53] Column (1) assigns the missing treated female the 25th percentile of each variable and the missing control female the 75th percentile of each variable (for all variables, higher percentiles correspond to "better" outcomes). Column (2) assigns the missing treated female the 10th percentile of each variable and the missing control female the 90th percentile of each variable. In both columns, the two significant outcomes - college attendance and marijuana use - remain significant.

Columns (3) through (6) address both follow-up and pre-treatment attrition. Column (3) assigns missing values as follows: the missing follow-up treated subject receives the 25th percentile for each variable, the missing follow-up control subject receives the 75th percentile for each variable, four of the missing pre-treatment treated subjects receive the 25th percentile for each variable, and two of the missing pre-treatment control subjects receive the 75th percentile for each variable.[54] Column (4) is identical to column (3) except that the missing follow-up subjects are assigned the 10th and 90th percentiles respectively. Column (5) is identical to column (4) except that the missing pre-treatment subjects are assigned the 10th and 90th percentiles respectively. Column (6) implements the "worst case" scenario: all attrition is assumed non-random, all missing subjects are assumed female unless otherwise identified, all missing treated subjects are assigned the 10th percentile values, and all missing control subjects are assigned the 90th percentile values. The worst case scenario assigns values to a total of seventeen missing subjects.

The results in columns (3) through (6) demonstrate that some Abecedarian effects retain significance under all but extreme assumptions about missing values. Both college attendance and marijuana use remain significant in columns (3) and (4). These variables lose significance in column (5), when six missing pre-treatment subjects are assigned values at the 10th and 90th percentiles of the distribution. The coefficients approach zero in column (6) under the worst case scenario; however, the assumptions underlying this scenario are implausible.

Murphy, 1999). Even if preschool induced a 50 percent change in death rates from these causes, total death rates would change by only 9.5 percent. The death rates in the actual sample match this prediction: two treatment group and two control group children died. Of course, it is theoretically possible that preschool could prevent some deaths and cause others, so dramatic effects for particular causes could be masked at the aggregate level. This seems unlikely.

[53] There is no high school graduation information for one additional treated female. Therefore, relative to the other results, the high school graduation results assign values for one additional treated female.

[54] A total of seven treated subjects and four control subjects are missing from pre-treatment attrition. I do not have information on their genders, so in the base case I assign genders to the missing pre-treatment subjects based on the gender distribution of the non-missing sample.

A.1.2 Perry

The Perry Preschool Project lost five children to pre-treatment attrition. Four treatment group children moved away before completing preschool, and one control group child died (Schweinhart, et al., 2005). None of these children entered the dataset. However, for several key measures, there is no follow-up attrition.

Table 12 presents estimates for key outcomes under three sets of assumptions. As with Abecedarian, the analysis focuses on females. The pre-treatment attrition in Perry is plausibly independent of treatment status. 80 percent of the pre-treatment attrition occurred when four treatment children moved away before completing the program. No control child moved away during the same period, and it is doubtful that the offer of free schooling would make a family *more* likely to leave the area. This attrition is therefore unlikely to be related to treatment status. An additional control child died at an early age and was not included in the sample. This death is unlikely to be the result of treatment status.[55]

Columns (1) and (2) in Table 12 address follow-up attrition only. Since marital status, high school graduation status, and government transfer data are available for all individuals, the reported coefficients for these variables are identical to the original results. Column (1) assigns missing treated subjects the 25th percentile of each variable conditional on high school graduation status. It assigns missing control subjects the 75th percentile of each variable conditional on high school graduation status.[56] All variables remain significant in column (1). Column (2) is identical to column (1) except that the 25th and 75th percentiles are replaced with the 10th and 90th percentiles respectively. Every variable except employment remains significant. Column (3) implements the "worst case" scenario. For variables with follow-up attrition, column (3) assigns missing treated subjects the 10th percentile of each variable conditional on high school graduation status and missing control subjects the 90th percentile of each variable conditional on high school graduation status. The four treated subjects that moved away are assumed to be female and assigned the 10th percentile of each variable. The one dead control subject is assumed to be female and assigned the 90th percentile of each variable. This worst case scenario eliminates the significance of most variables. However, the high school graduation effect remains significant despite the extreme assumptions underlying this scenario.

[55] Please see the note in Section A.1.1 regarding attrition due to death.

[56] This procedure leverages information contained in the complete high school graduation data for predictive purposes.

A.2 Violation of Random Assignment

For the most part, families complied with their initial group assignments. Those that refused were generally dropped from the data, as described in Section A.1. However, in the Perry Preschool Project, several children with working mothers were exchanged with select control group children. Two of these switches may have occurred without replacement. The exchanges were made because the employed mothers could not accommodate the program's weekly home visits. Replacement children were purportedly matched on initial IQ, but confounding may still occur because maternal presence at an early age could affect later outcomes.

In no case did the Perry researchers record original assignment status. This fact precludes the use of an instrumental variables approach, so I perform alternative tests to gauge the impact of these violations. First, I condition on initial maternal employment status. However, five children with employed mothers were not transfered from the treatment group. Conditioning upon maternal employment status is therefore insufficient, because children with employed mothers who switched may differ in important ways from those with employed mothers who stayed. Furthermore, conditioning on maternal employment does not account for the control children who were exchanged to the treatment group. If these children were matched with the maternal employment children, they may differ from the average child in expectation. To address these issues, I examine a range of possible group assignments and the corresponding coefficient estimates.

The first two columns of Table 13 present results for key outcomes for both genders controlling for the effect of maternal employment at entry. These results do not differ markedly from the original estimates. In fact, the coefficients are of slightly greater magnitude after controlling for maternal employment. All female effects remain significant, and one male effect - lifetime arrests at 27 - achieves marginal significance.

I conduct further analysis for key female outcomes under the assumption that four treatment children with employed mothers were switched with four control children without employed mothers. Records indicate that either two or five children with employed mothers switched from the treatment group, but probability estimates indicate that the number could have been as high as eight (Schweinhart, et al., 2005). There is no record of the gender distribution of exchanges. However, the data suggest that approximately three females switched from the treatment group, since there are nine control females with employed mothers as compared to three treated females with employed mothers. The assumption that four treated females were exchanged is therefore likely to overestimate the total number of female exchanges.

The exchange analysis examines every possible combination of switches that swaps four treated females

with employed mothers for four control females without employed mothers. The number of possible combinations totals 921,600. For each combination, I estimate the treatment effect for six key outcomes using instrumental variables. The hypothesized original group assignment serves as the instrument.

Because each individual carries an entire set of outcomes, it is meaningless to tabulate the resulting t-statistics in isolation. In order to compare different combinations of exchanges, I construct an average t-statistic for each combination equal to the mean of the six estimated t-statistics.[57] I then rank combinations according to their average t-statistics.

The last five columns of Table 13 report female results for different quantiles of the average t-statistic. The first column presents results at the median of the distribution, the second at the 25th quantile, the third at the 10th quantile, the fourth at the 1st quantile, and the fifth at the distribution's minimum value. At the median, the coefficients are of similar magnitude to the original OLS results, but the standard errors have increased because the instrument is not perfectly correlated with treatment status. Consequently, some results are insignificant, but the two arrest variables and the graduation variable remain significant. At the 10th quantile, the graduation and arrest outcomes attain marginal significance. At the bottom of the distribution they are all insignificant. However, these quantiles are identified ex post. When the Perry researchers chose which control individuals to exchange with treated individuals, they could only guess at future outcomes. Even if the researchers tried make exchanges that would benefit the treatment group and hurt the control group (an implausible assumption), it is unlikely they could achieve that goal to as great a degree as implied in the 10th or 1st quantile columns. The last three columns of Table 13 therefore correspond to very extreme assumptions, but even in the 1st quantile column one coefficient remains statistically significant. It is therefore unlikely that exchanges based on maternal employment status drive the significance of the results.

A.3 Clustering

The p-values presented in Section (4) are robust to distributional assumptions.[58] Nevertheless, clustering issues could bias the standard errors, causing conventional tests to overstate the significance of the results, particularly in the case of the Perry Preschool Project.

It is well established that clustering - or correlation across observations - can severely inflate test statistics

[57] When constructing the average t-statistic, I reverse the sign on the two arrest t-statistics and the government transfer t-statistic, so that positive t-statistics always correspond to "better" outcomes.

[58] These p-values do not differ markedly from the p-values generated by conventional t-tests, so standard OLS t-tests are presented for the remaining results.

if not properly accounted for (Bertrand, Duflo, and Mullainathan, 2004). In these experiments, there are two possible sources of interdependence that could be correlated within treatment status groups.[59] First, peer or class effects might lead to correlations between students within a given preschool class.[60] Second, the automatic assignment of younger siblings to the same treatment group as their older siblings reduces the number of independent observations.

Previous research has demonstrated that negative peer effects can lower class achievement (for example, Figlio, 2005). It is therefore plausible that a poorly behaved child may reduce the performance of her preschool peers, implying an intra-class correlation.[61] Furthermore, within each class the treatment variable is perfectly correlated. The standard errors could therefore be too small because children within each class are mistakenly treated as independent observations.

To address this problem, I collapse the data down to cohort-by-treatment status means. For the Perry females, with five cohorts and two treatment statuses, this procedure reduces the dataset to ten observations. I estimate an OLS regression using these ten observations. The first row of Table 14 presents the results. Despite the small sample, five of the six key variables remain statistically significant.[62] The only outcome that loses significance is the employment variable. I cannot run a similar regression for the Abecedarian children as I do not have their cohort identification data, but the Perry analysis suggest that intra-class clustering does not drive the significance of the results.

Another clustering problem arises from the assignment of younger siblings to the same treatment status as their older siblings. Performance is almost surely correlated within families, and this assignment mechanism guarantees that treatment status is also correlated within families. To address this problem, I restrict the sample to eldest siblings and only children. For Perry females, this restriction decreases the sample size from 51 to 37. The results for Perry females are reported in the second row of Table 14. All presented outcomes remain strongly significant. In the Abecedarian program, the sample contains only two sibling pairs, so an older sibling analysis is unnecessary.

[59] Inflation only occurs when there is a similar correlation structure in both the dependent and independent variables. For example, the fact that cohorts might face similar shocks will not bias the standard errors because treatment status is randomly assigned within a given cohort.

[60] Angrist and Lang (2004), for example, find evidence of negative peer effects in the Boston METCO program.

[61] Peer effects may operate more strongly for poorly behaved children than for well behaved children. In that case, they will tend to reduce the estimated effect. However, this is not a source of coefficient bias, since it is a direct consequence of the preschool program. Rather, it only effects the external validity of the results when applied to different demographic groups.

[62] Donald and Lang (2001) argue that t-statistics can be misleading when the number of groups - in this case, cohort-by-treatment status units - is small, because common shocks may not be normally distributed. However, for these results I conducted simulations drawing common group effects from a distribution with all its probability weight at the tails. These simulations did not generate poorly distributed t-statistics.

A final clustering issue is the possibility of teacher effects.[63] Individual level data on teacher assignment is not available. However, Perry Preschool employed ten teachers, the Early Training Project employed two teachers and several assistants, and the Abecedarian Project employed multiple teachers (the exact number is unclear). It is therefore unlikely that the observed effects are the result of one or two stellar teachers.

References

Abadie, Alberto, Joshua Angrist, and Guido Imbens (2002) 'Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings.' *Econometrica* 70(1), 91–117

American Psychiatric Association (1994) *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)* (American Psychiatric Association)

Angrist, Joshua, and Kevin Lang (2004) 'Does School Integration Generate Peer Effects? Evidence from Boston's Metco Program.' *American Economic Review* 94(5), 1613–1634

Angrist, Joshua, Guido Imbens, and Donald Rubin (1996) 'Identification of Causal Effects Using Instrumental Variables.' *Journal of the American Statistical Association* 91(434), 444–455

Australian Institute of Family Studies (2005) 'Growing Up in Australia: The Longitudinal Study of Australian Children: 2004 Annual Report.' Australian Institute of Family Studies

Berreuta-Clement, J., L. Schweinhart, W. S. Barnett, A. Epstein, and D. Weikart (1984) *Changed Lives: The Effects of the Perry Preschool Program on Youths Through Age 19* (High/Scope Press)

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004) 'How Much Should We Trust Difference in Difference Estimates?' *Quarterly Journal of Economics* 119(1), 249–275

Campbell, Frances, and Craig 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), 684–698

Campbell, Frances, Craig Ramey, Elizabeth Pungello, Joseph Sparling, and Shari Miller-Johnson (2002) 'Early Childhood Education: Young Adult Outcomes From the Abecedarian Project.' *Applied Developmental Science* 6(1), 42–57

Campbell, Frances, Elizabeth Pungello, Shari Miller-Johnson, Margaret Burchinal, and Craig Ramey

[63] Technically, teacher effects would be an issue of external validity, not internal validity.

(2001) 'The Development of Cognitive and Academic Abilities: Growth Curves From an Early Childhood Educational Experiment.' *Developmental Psychology* 37(2), 231–242

Carneiro, Pedro, and James Heckman (2003) 'Human Capital Policy.' In *Inequality in America: What Role for Human Capital Policies?,* ed. James Heckman and Alan Krueger (MIT Press)

Clarke, Stevens, and Frances Campbell (1998) 'Can Intervention Early Prevent Crime Later? The Abecedarian Project Compared with Other Programs.' *Early Childhood Research Quarterly* 13(2), 319–343

Cunha, Flavio, James Heckman, Lance Lochner, and Dimitriy Masterov (2005) 'Interpreting the Evidence on Life Cycle Skill Formation.' NBER Working Paper Series, Working Paper 11331

Currie, Janet (2001) 'Early Childhood Education Programs.' *Journal of Economic Perspectives* 15(2), 213–238

Currie, Janet, and Duncan Thomas (1995) 'Does Head Start Make a Difference?' *American Economic Review* 85(3), 341–364

— (2000) 'School Quality and the Longer-Term Effects of Head Start.' *Journal of Human Resources* 35(4), 755–774

Currie, Janet, and Matthew Neidell (2004) 'Getting Inside the "Black Box" of Head Start Quality: What Matters and What Doesn't.' UCLA Department of Economics, manuscript

Dee, Thomas (2005) 'A Teacher Like Me: Does Race, Ethnicity or Gender Matter?' *American Economic Review Papers and Proceedings* 95(2), 158–165

Donald, Stephen, and Kevin Lang (2004) 'Inference with Difference in Differences and Other Panel Data.' Boston University Department of Economics, manuscript

Figlio, David (2005) 'Boys Named Sue: Disruptive Children and their Peers.' National Bureau of Economic Research Working Paper No. 11277

Fredriksson, Peter, and Björn Öckert (2005) 'Is Early Learning Really More Productive? The Effect of School Starting Age on School and Labor Market Performance.' IZA Discussion Paper Series, No. 1659

Garces, Eliana, Duncan Thomas, and Janet Currie (2002) 'Longer-Term Effects of Head Start.' *American Economic Review* 92(4), 999–1012

Gray, Susan, and Rupert Klaus (1965) 'An Experimental Preschool Program for Culturally Deprived Children.' *Child Development* 36(4), 887–898

— (1970) 'The Early Training Project: A Seventh-Year Report.' *Child Development* 41(4), 909–924

Gray, Susan, Barbara Ramsey, and Rupert Klaus (1982) *From 3 to 20: The Early Training Project* (University Park Press)

Hanushek, Eric (1996) 'School Resources and Student Performance.' In *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success.,* ed. Gary Burtless (Brookings Institution)

— (2003) 'Comments.' In *Inequality in America: What Role for Human Capital Policies?,* ed. James Heckman and Alan Krueger (MIT Press)

Heckman, James (2005) 'Invited Comments.' In *Lifetime Effects: The High/Scope Perry Preschool Study Through Age 40,* ed. L. Schweinhart, et al. (High/Scope Press)

Heckman, James, and Yonah Rubinstein (2001) 'The Importance of Noncognitive Skills: Lessons from the GED Testing Program.' *American Economic Review* 91(2), 145–149

Horowitz, Joel (2001) 'The Bootstrap in Econometrics.' In *Handbook of Econometrics,* ed. James Heckman and Edward Leamer, vol. 5 (Elsevier Science B.V.) chapter 52, pp. 3159–3228

Hoyert, Donna, Kenneth Kochanek, and Sherry Murphy (1999) 'Deaths: Final Data for 1997.' *National Vital Statistics Reports*

Kirp, David (2005) 'All My Children.' *The New York Times,* July 31, 2005, Section 4A, 20

Kling, Jeffrey, and Jeffrey Liebman (2004) 'Experimental Analysis of Neighborhood Effects on Youth.' Kennedy School of Government Working Paper No. RWP04-034

Krueger, Alan (1999) 'Experimental Estimates of Education Production Functions.' *The Quarterly Journal of Economics* 114(2), 497–532

— (2003) 'Inequality, Too Much of a Good Thing.' In *Inequality in America: What Role for Human Capital Policies?,* ed. James Heckman and Alan Krueger (MIT Press)

Miller, Jerome (1992) 'Hobbling a Generation: Young African American Males in the Criminal Justice System of America's Cities: Baltimore, Maryland.' National Center on Institutions and Alternatives

O'Brien, Peter (1984) 'Procedures for Comparing Samples with Multiple Endpoints.' *Biometrics* 40(4), 1079–1087

Oden, S., L. Schweinhart, D. Weikart, S. Marcus, and Y. Xie (2000) *Into Adulthood: A Study of the Effects of Head Start* (High/Scope)

Ramey, Craig, Keith Yeates, and Elizabeth Short (1984) 'The Plasticity of Intellectual Development: Insights from Preventative Intervention.' *Child Development* 55(5), 1913–1925

Schweinhart, L., H. Barnes, and D. Weikart (1993) *Significant Benefits: The High/Scope Perry Preschool*

Study Through Age 27 (High/Scope Press)

Schweinhart, L., J. Montie, Z. Xiang, W. S. Barnett, C. Belfield, and M. Nores (2005) *Lifetime Effects: The High/Scope Perry Preschool Study Through Age 40* (High/Scope Press)

Simon, Julian (1997) *Resampling: The New Statistics* (Resampling Stats)

Weikart, D., D. Deloria, S. Lawser, and R. Wiegerink (1970) *Longitudinal Results of the Ypsilanti Perry Preschool Project* (High/Scope Press)

Westfall, Peter, and S. Young (1993) *Resampling-Based Multiple Testing* (John Wiley and Sons)

Yucesan, Enver (1995) 'Using Nonparametric Statistics in Simulation Analysis: A Review.' In *Proceedings of the 1995 Winter Simulation Conference,* ed. C. Alexopoulos, K. Kang, W. Lilegdon, and D. Goldsman pp. 141–146

Figure 1: Effects of Preschool on Teen and Adult Outcomes

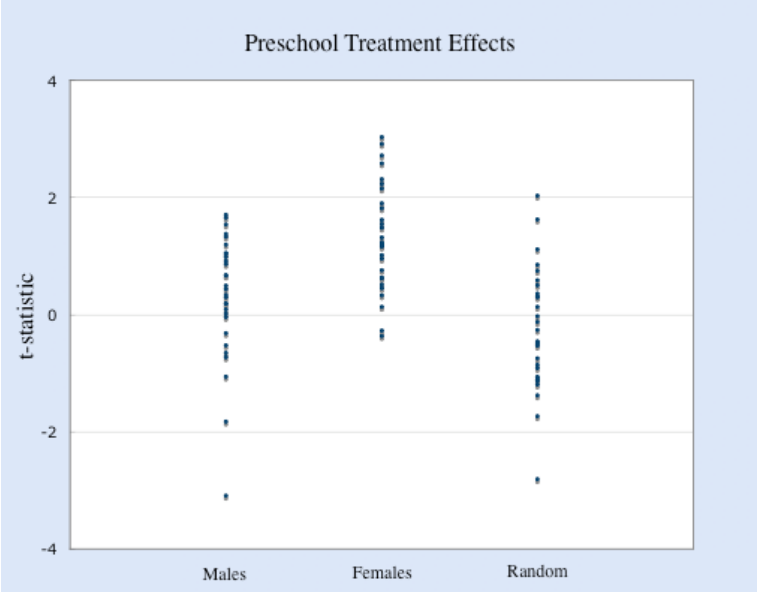


Table 1: Summary Statistics

Variable	Abecedarian	Perry	Early Training
Percent treated	51.4	47.2	67.7
	(50.2)	(50.1)	(47.1)
Percent female	53.2	41.5	46.2
	(50.1)	(49.5)	(50.2)
IQ age 5	97.8	88.9	91.5
	(12.6)	(12.9)	(13.6)
IQ age 14-17	93.2	80.9	77.7
	(10.3)	(11.0)	(13.2)
Percent retained in grade	45.6	37.5	54.2
	(50.1)	(48.6)	(50.2)
Percent graduate HS	69.9	61.8	60.0
	(46.1)	(48.8)	(49.4)
Percent employed as adult	57.3	62.1	N/A
	(49.7)	(48.7)	
Percent with criminal record	43.3	52.8	N/A
	(49.8)	(50.1)	

Notes: Parentheses contain standard deviations.

Table 2: Effects on Pre-Teen IQ Scores

			Female				Male				Gender Interaction
Outcome	Age	Project	**Effect**	CM	*p*-val	N	**Effect**	CM	*p*-val	N	*t*-stat
IQ	5	ABC	4.94	96.76	0.182	48	10.19	90.81	0.005	47	-1.05
			(3.58)				(3.52)				
IQ	6.5	ABC	5.13	92.96	0.135	46	7.18	92.10	0.058	45	-0.41
			(3.35)				(3.65)				
IQ	12	ABC	8.35	87.35	0.004	52	3.21	90.48	0.291	49	1.24
			(2.75)				(3.10)				
IQ	5	Perry	12.67	81.65	0.004	39	10.61	84.79	0.000	54	0.40
			(4.30)				(2.84)				
IQ	6	Perry	3.75	87.16	0.243	48	5.66	85.82	0.037	72	-0.46
			(3.21)				(2.68)				
IQ	10	Perry	4.96	81.79	0.169	43	-2.33	86.03	0.375	71	1.70
			(3.45)				(2.56)				
IQ	5	ETP	13.55	87.60	0.018	30	4.43	87.18	0.232	34	1.28
			(6.09)				(3.75)				
IQ	7	ETP	8.61	89.89	0.119	29	4.11	92.89	0.346	30	0.57
			(6.69)				(4.25)				
IQ	10	ETP	9.79	81.56	0.069	29	-3.17	88.33	0.505	27	1.68
			(5.73)				(5.15)				

Notes: Parentheses contain robust standard errors. CM refers to control mean. Sample size varies within experiments due to attrition for some variables. *P*-values are computed as described in Section (3); *t*-statistics test the difference between female and male treatment effects.

Table 3: Effects on Pre-Teen Primary School Outcomes

			Female				Male				Gender Interaction
Outcome	Age	Project	**Effect**	CM	*p*-val	N	**Effect**	CM	*p*-val	N	*t*-stat
Retained	12	ABC	-0.229	0.429	0.082	53	-0.188	0.545	0.201	50	-0.21
			(0.125)				(0.142)				
Spec Educ	12	ABC	-0.066	0.296	0.576	53	-0.269	0.591	0.054	50	1.10
			(0.123)				(0.140)				
Repeat Grade	12	Perry	-0.201	0.409	0.134	46	0.078	0.389	0.514	66	-1.51
			(0.137)				(0.124)				
Spec Educ	17	Perry	-0.262	0.462	0.060	51	-0.037	0.462	0.741	72	-1.28
			(0.129)				(0.119)				
Retained	17	ETP	-0.284	0.600	0.156	29	0.100	0.600	0.514	30	-1.40
			(0.195)				(0.192)				
Special Help	17	ETP	0.116	0.200	0.529	29	0.036	0.364	0.832	31	0.31
			(0.171)				(0.188)				

Notes: Parentheses contain robust standard errors. CM refers to control mean. Sample size varies within experiments due to attrition for some variables. *P*-values are computed as described in Section (3); *t*-statistics test the difference between female and male treatment effects.

Table 4: Summary Index Effects

		Female			Male			Gender Interaction
Project	Age	**Effect**	FWE p-val	N	**Effect**	FWE p-val	N	t-stat
ABC	Pre-Teen	0.445	0.117	54	0.417	0.187	51	0.11
		(0.194)			(0.181)			
Perry	Pre-Teen	0.537	0.026	51	0.150	0.941	72	1.53
		(0.177)			(0.172)			
ETP	Pre-Teen	0.380	0.346	30	0.142	0.960	34	0.67
		(0.270)			(0.238)			
ABC	Teen	0.422	0.153	53	0.162	0.941	51	0.93
		(0.202)			(0.194)			
Perry	Teen	0.613	0.003	51	0.035	0.976	72	3.32
		(0.156)			(0.096)			
ETP	Teen	0.551	0.346	29	0.097	0.976	32	0.95
		(0.327)			(0.345)			
ABC	Adult	0.452	0.022	53	0.312	0.369	51	0.64
		(0.144)			(0.166)			
Perry	Adult	0.358	0.117	51	-0.017	0.976	72	1.88
		(0.151)			(0.130)			
ETP	Adult	-0.067	0.709	29	-0.654	0.090	31	1.82
		(0.188)			(0.257)			

Notes: Parentheses contain OLS standard errors. FWE p-values are computed as described in Section (3); t-statistics test the difference between female and male treatment effects. See Table 5 for the components of each summary index.

Table 5: Summary Index Components

Project	Stage	Summary Index Components
ABC	Pre-Teen	IQ (5, 6.5, 12), Retained in Grade (12), Special Education (12)
Perry	Pre-Teen	IQ (5, 6, 10), Repeat Grade (17), Special Education (17)
ETP	Pre-Teen	IQ (5, 7, 10), Retained in Grade (17), Special Help (17)
ABC	Teen	IQ (15), HS Grad (18), Teen Parent (19)
Perry	Teen	IQ (14), HS Grad (18), Unemployed (19), Transfers (19), Teen Parent (19), Arrested (19)
ETP	Teen	IQ (17), HS Drop Out (18), Worked (18)
ABC	Adult	College (21), Employed (21), Convicted (21), Felon (21), Jailed (21), Marijuana (21)
Perry	Adult	College (27), Employed (27, 40), Income (27, 40), Criminal Record (27), Arrests (27), Drugs (27), Married (27)
ETP	Adult	College (21), Receive Income (21), On Welfare (21)

Notes: Age of measurement in parentheses.

Table 6: Effects on Teenage Academic Outcomes

			Female				Male				Gender Interaction
Outcome	Age	Project	**Effect**	CM	*p*-val	N	**Effect**	CM	*p*-val	N	*t*-stat
IQ	15	ABC	4.22	89.50	0.142	53	4.66	92.48	0.091	51	-0.11
			(2.85)				(2.79)				
IQ	14	Perry	2.64	76.77	0.313	46	-0.96	83.26	0.761	64	0.91
			(2.57)				(3.03)				
IQ	17	ETP	2.08	76.11	0.744	25	1.64	76.78	0.737	28	0.05
			(6.80)				(5.09)				
HS Grad	18	ABC	0.226	0.607	0.086	52	-0.096	0.739	0.465	51	1.80
			(0.122)				(0.131)				
HS Grad	18	Perry	0.494	0.346	0.000	51	-0.061	0.667	0.583	72	3.32
			(0.121)				(0.115)				
Ever Drop	18	ETP	-0.289	0.500	0.107	29	-0.095	0.545	0.676	31	-0.72
Out of HS			(0.190)				(0.193)				

Notes: Parentheses contain robust standard errors. CM refers to control mean. Sample size varies within experiments due to attrition for some variables. *P*-values are computed as described in Section (3); *t*-statistics test the difference between female and male treatment effects.

Table 7: Effects on Teenage Economic and Social Outcomes

			Female				Male				Gender Interaction
Outcome	Age	Project	**Effect**	CM	*p*-val	N	**Effect**	CM	*p*-val	N	*t*-stat
Unemp	19	Perry	-0.308	0.708	0.028	49	-0.021	0.385	0.877	72	-1.60
			(0.138)				(0.116)				
Transfers	19	Perry	-1,569	2,828	0.035	51	-28	398	0.933	72	-1.96
			(722)				(319)				
Ever Work	18	ETP	0.125	0.500	0.581	22	-0.063	1.000	0.641	23	0.73
			(0.249)				(0.063)				
Teen Parent	19	ABC	-0.211	0.571	0.133	53	-0.126	0.304	0.315	51	-0.47
			(0.137)				(0.123)				
Had Child	19	Perry	-0.187	0.667	0.209	49	-0.044	0.256	0.666	72	-0.82
			(0.142)				(0.101)				
Arrested	19	Perry	-0.337	0.417	0.006	49	-0.079	0.564	0.527	72	-1.54
			(0.117)				(0.119)				

Notes: Parentheses contain robust standard errors. CM refers to control mean. Sample size varies within experiments due to attrition for some variables. *P*-values are computed as described in Section (3); *t*-statistics test the difference between female and male treatment effects.

Table 8: Effects on Adult Academic Outcomes

			Female				Male				Gender Interaction
Outcome	Age	Project	**Effect**	CM	*p*-val	N	**Effect**	CM	*p*-val	N	*t*-stat
In College	21	ABC	0.293	0.107	0.015	53	0.148	0.174	0.258	51	0.87
			(0.116)				(0.121)				
Any College	27	Perry	0.160	0.280	0.256	50	-0.005	0.308	0.978	72	0.94
			(0.137)				(0.110)				
In Post HS Educ	21	ETP	0.121	0.300	0.537	29	-0.486	0.636	0.005	31	2.37
			(0.191)				(0.171)				

Notes: Parentheses contain robust standard errors. CM refers to control mean. Sample size varies within experiments due to attrition for some variables. *P*-values are computed as described in Section (3); *t*-statistics test the difference between female and male treatment effects.

Table 9: Effects on Adult Economic Outcomes

			Female				Male				Gender Interaction
Outcome	Age	Project	**Effect**	CM	*p*-val	N	**Effect**	CM	*p*-val	N	*t*-stat
Employed	21	ABC	0.104	0.536	0.432	53	0.188	0.455	0.196	50	-0.43
			(0.137)				(0.142)				
Employed	27	Perry	0.255	0.545	0.076	47	0.036	0.564	0.765	69	1.20
			(0.136)				(0.121)				
Annual Income	27	Perry	2,567	8,986	0.353	47	2,363	12,495	0.382	66	0.05
			(2,686)				(2,708)				
Employed	40	Perry	0.015	0.818	0.922	46	0.200	0.500	0.109	66	-1.12
			(0.115)				(0.120)				
Annual Income	40	Perry	3,492	17,374	0.536	46	6,228	21,119	0.302	66	-0.34
			(5,491)				(5,958)				
Receive Income	21	ETP	-0.074	0.600	0.688	29	-0.159	0.909	0.303	31	0.36
			(0.200)				(0.134)				
Receive Welfare	21	ETP	-0.042	0.200	0.805	30	N/A	0.000	N/A	35	N/A
			(0.157)				(N/A)				

Notes: Parentheses contain robust standard errors. CM refers to control mean. Sample size varies within experiments due to attrition for some variables. *P*-values are computed as described in Section (3); *t*-statistics test the difference between female and male treatment effects.

Table 10: Effects on Adult Social Outcomes

			Female				Male				Gender Interaction
Outcome	Age	Project	**Effect**	CM	*p*-val	N	**Effect**	CM	*p*-val	N	*t*-stat
Convicted	21	ABC	-0.101	0.143	0.224	52	-0.089	0.348	0.523	50	-0.08
			(0.079)				(0.133)				
Felony	21	ABC	N/A	0.000	N/A	52	-0.113	0.261	0.369	50	N/A
			N/A				(0.117)				
Jailed	21	ABC	-0.030	0.071	0.703	52	-0.177	0.391	0.160	51	1.01
			(0.065)				(0.131)				
Marijuana User	21	ABC	-0.317	0.357	0.003	53	-0.127	0.435	0.390	49	-1.10
			(0.101)				(0.140)				
Criminal Record	27	Perry	-0.146	0.346	0.260	51	-0.021	0.718	0.824	72	-0.75
			(0.125)				(0.109)				
Lifetime Arrests	27	Perry	-1.95	2.27	0.012	49	-2.31	6.10	0.133	72	0.21
			(0.83)				(1.50)				
Ever Used Drugs	27	Perry	-0.157	0.300	0.213	41	0.198	0.189	0.073	68	-2.08
			(0.131)				(0.110)				
Married	27	Perry	0.317	0.083	0.008	49	0.002	0.256	0.983	70	2.01
			(0.115)				(0.107)				

Notes: Parentheses contain robust standard errors. CM refers to control mean. Sample size varies within experiments due to attrition for some variables. *P*-values are computed as described in Section (3); *t*-statistics test the difference between female and male treatment effects.

Appendix Tables

Table 11: Attrition Analysis for Key Abecedarian Variables

Outcome	Age	(1)	(2)	(3)	(4)	(5)	(6)
High School Grad	18	0.149	0.149	0.022	0.022	0.022	-0.061
		(0.125)	(0.125)	(0.124)	(0.124)	(0.124)	(0.120)
Attending College	21	0.281	0.247	0.237	0.204	0.140	0.061
		(0.110)	(0.115)	(0.102)	(0.106)	(0.113)	(0.111)
Marijuana User	21	-0.306	-0.268	-0.289	-0.256	-0.123	-0.030
		(0.102)	(0.108)	(0.093)	(0.098)	(0.113)	(0.113)
Teen Parent	19	-0.167	-0.167	-0.049	-0.049	-0.049	0.030
		(0.135)	(0.135)	(0.130)	(0.130)	(0.130)	(0.125)

Notes: Parentheses contain OLS standard errors.

Table 12: Attrition Analysis for Key Perry Variables

Outcome	Age	(1)	(2)	(3)
Employed	27	0.300	0.185	0.071
		(0.130)	(0.128)	(0.127)
Married	27	0.285	0.285	0.179
		(0.118)	(0.118)	(0.118)
Ever Arrested	19	-0.305	-0.305	-0.179
		(0.113)	(0.113)	(0.118)
High School Grad	18	0.494	0.494	0.357
		(0.122)	(0.122)	(0.126)
Transfers	19	-1569	-1569	-945
		(729)	(729)	(716)
Lifetime Arrests	27	-1.95	-1.95	-1.39
		(0.84)	(0.84)	(0.81)

Notes: Parentheses contain OLS standard errors.

Table 13: Effects of Maternal Employment on Key Perry Results

		Control for WM		Alternative Assumptions on WM Swaps				
Outcome	Age	Female	Male	50th	25th	10th	1st	Lowest
Employed	27	0.316	0.115	0.225	0.244	0.048	-0.003	-0.003
		(0.141)	(0.122)	(0.274)	(0.234)	(0.260)	(0.243)	(0.243)
Married	27	0.318	0.039	0.306	0.192	0.391	0.259	-0.017
		(0.123)	(0.111)	(0.226)	(0.199)	(0.213)	(0.198)	(0.214)
Ever Arrested	19	-0.398	-0.083	-0.661	-0.500	-0.352	-0.431	-0.293
		(0.118)	(0.126)	(0.242)	(0.199)	(0.210)	(0.197)	(0.195)
High School Grad	18	0.530	-0.005	0.581	0.530	0.373	0.296	0.296
		(0.126)	(0.119)	(0.228)	(0.200)	(0.216)	(0.206)	(0.206)
Transfers	19	-1765	-144	-2254	-1078	-2102	-794	-670
		(756)	(337)	(1368)	(1200)	(1287)	(1213)	(1217)
Lifetime Arrests	27	-2.30	-2.90	-3.70	-2.77	-2.78	-2.30	-2.17
		(0.86)	(1.56)	(1.64)	(1.40)	(1.50)	(1.39)	(1.39)

Notes: WM = Working Mothers. Results under alternative assumptions are estimated using hypothesized group assignment as an instrument for treatment status. Parentheses contain OLS standard errors when controlling for working mothers, and IV standard errors when examining results under alternative assumptions about the working mother swaps. Sample size varies within columns due to attrition for some variables.

Table 14: Effects of Clustering on Key Perry Results

Model	**Employed at 27**	**Married at 27**	**Arrested by 19**	**High School Graduate**	**Transfers at 19**	**Lifetime Arrests at 27**
Collapsed to Cohort	0.195	0.336	-0.303	0.477	-1703	-1.59
by Treatment Means	(0.178)	(0.171)	(0.136)	(0.166)	(889)	(0.74)
Eldest Siblings and	0.342	0.409	-0.307	0.561	-2563	-1.85
Only Children Sample	(0.151)	(0.146)	(0.139)	(0.134)	(859)	(0.88)

Notes: For results estimated using cohort by treatment means, N=10. Parentheses contain OLS standard errors.

B Stata Pseudo-Code

Sample Stata code for the free step-down resampling algorithm follows. Some code has been changed to improve readability and would not run as literally written. This code is adapted from Algorithm 2.8 in Westfall and Young (1993).

```
local counter = 1
* run the original regressions and create the p-val simulation storage counters
foreach lhsvar in outcome-varlist {
regress ‘lhsvar’ treated
replace t-stat = abs(_b[treated]/_se[treated]) in ‘counter’
replace p-val = 2*ttail(e(N),t-stat)
local ‘lhsvar’-count = 0
local counter = ‘counter’ + 1
}

* sort the regressions according to ascending p-value.
sort p-val
sort outcome-varlist by ascending p-val

* store the total number of tests originally conducted
local endvar = ‘counter’ - 1
* initialize the simulation counter
local iteration = 1

* run 10,000 iterations of the simulation; record results in p-val storage counters
while ‘iteration’ ¡= 10000 {
replace simtreatment-uni = uniform()
replace simtreatment = (simtreatment-uni > 0.5) if perry==1 or abc==1
replace simtreatment = (simtreatment-uni > 0.67) if etp==1
```

```
local counter = 1
foreach lhsvar of outcome-varlist {
regress ‘lhsvar’ simtreatment
replace t-stat-sim = abs(_b[simtreatment]/_se[simtreatment]) in ‘counter’
replace p-val-sim = 2*ttail(e(N),t-stat-sim) in ‘counter’
local counter = ‘counter’ + 1
}
* enforce monotonicity in the simulated p-vals and then tabulate whether the simulated p-vals are less than the respective original p-vals
local countdown = ‘endvar’
foreach lhsvar of reverse-outcome-varlist {
replace p-val-sim = min(p-val-sim, p-val-sim[_n+1]) in ‘countdown’
if p-val-sim[‘countdown’] <= p-val[‘countdown’] {
local ‘lhsvar’-count = ‘lhsvar’-count + 1
}
local countdown = ‘countdown’ - 1
}
local iteration = ‘iteration’ + 1
}

* calculate the adjusted p-val as the ratio of the number of times that the simulated p-vals are less than the original p-val divided by the total number of iterations; enforce the ordering of the original p-values
local counter = 1
foreach lhsvar of outcome-varlist {
replace p-vals = max(round(‘lhsvar’-count/10000, .001), p-vals[‘counter’-1]) in ‘counter’
local counter = ‘counter’ + 1
}
```