Multiple Inference and Gender Differences in the Effects of Preschool: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects [*]

Michael Anderson
Department of Agricultural and Resource Economics, U.C. Berkeley

July 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 two core issues that have received little attention in previous analyses: treatment effect heterogeneity by gender and over-rejection of the null hypothesis due to multiple inference. The primary finding of this reanalysis is that girls garnered substantial short- and long-term benefits from the interventions. However, there were no significant long-term benefits for boys. These conclusions would not be apparent when using ”naive” estimators that do not adjust for multiple inference.

[*] Michael Anderson is Assistant Professor, Department of Agricultural and Resource Economics, University of California, Berkeley, CA 94720 (E-mail: milypan@alum.mit.edu). Funding from the National Institute on Aging, through Grant Number T32-AG00186 to the National Bureau of Economic Research, is gratefully acknowledged. The author thanks his advisors, Joshua Angrist, David Autor, and Jon Gruber, for their insightful comments and suggestions. He also thanks Daron Acemoglu, Bruce Meyer, Michael Wolf, and the participants of the MIT Public/Labor Seminar and MIT Labor Lunch for helpful input. He is 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].

1 Introduction

The education literature contains dozens of papers showing inconsistent or low returns to publicly funded human capital investments (cf. Hanushek, 1986; Slavin, 1989; Zellman, et al., 1998; Stecher, McCaffrey, and Bugliari, 2003). 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 often cited as proof of efficacy for many types of early interventions (cf. Currie, 2001; Cunha, et al., 2005). The experiments underlie the growing movement for universal pre-kindergarten education (Kirp, 2005) and 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 the 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, overcoming the problem of confounding that affects many observational studies. Following initial assignment to treatment and control groups, treated children in each experiment received several years of preschool education (intensity differed across programs). Intervention continued until the children began regular schooling. At that point, further intervention was limited to data collection; children in both treatment and control groups received a series of standardized tests lasting through their teenage years. Researchers also conducted subject interviews and examined school and 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 the Abecedarian and Perry experiments, attrition materialized before preschool treatment and during the collection of follow-up data. As a result, the 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. More importantly, 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. This problem is well known in the theoretical literature (cf. Romano and Wolf, 2005) and the biostatistics field (cf. Hochberg, 1988), but it has yet to receive significant attention in the policy evaluation literature. 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 results (cf. Currie and Thomas, 1995; Krueger, 2003).

This paper has two related objectives. First, it implements a unified statistical framework to directly address concerns about sample size and multiple inference. This general framework is easily applicable to many types of program evaluation studies. Second, in recognition of the emerging female-male scholastic achievement gap (Lewin, 2006), the paper simultaneously examines all three studies to estimate the long-term effects of preschool separately for both males and females.[1] The organization is 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. 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, particularly academic achievement. However, treatment effects are minimal or nonexistent for males - a fact that would not be clear using ”naive” analyses that fail to account for multiple inference.

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. The

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

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. In contrast to the other programs, Abecedarian control children received minor interventions: iron fortified formula, free diapers, and supportive social services when appropriate. Of the three preschool projects, Abecedarian was the most intensive (for further details, see Campbell and Ramey, 1994).

The Abecedarian dataset contains 111 children; 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. The data come 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. 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.

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. Treated children entered the program at age three and remained in it for two years. The program implemented the ideas of Jean Piaget and focused on language skills, socialization, numbers, space, and time. Treated children attended the program five mornings per week from October through May and received one 90 minute home visit per week (for further details, see Schweinhart, et al., 2005).

The Perry dataset contains 123 individuals, 58 in the treatment group and 65 in the control group. 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 program 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 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.

The Early Training Project gathered data on 88 children. However, 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 randomly 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. Since the children in the distal control group were not randomly assigned, and their observable characteristics are not similar to the local control group (Anderson, 2006), I drop them from the analysis. This choice results in a total sample of 65: 44 treated children and 21 control children.

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.

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 77.7 to 93.2. 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 disadvantaged neighborhoods (cf. Miller, 1992).

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 estimates 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."[2]

To conduct inference, I compute Huber-White standard errors that are robust to heteroskedasticity (White, 1980). 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 the underlying data is distributed non-normally in some cases (Horowitz, 2001). 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 (Efron

[2] 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.

and Tibshirani, 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. 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. Formally, this procedure tests the sharp null hypothesis of no treatment effect, so rejection implies that the treatment has some distributional effect.

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. However, it 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. 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 - the randomization arises from the experimental design (cf. Fisher, 1935). It samples the preschool assignments z_i^* with replacement because the original assignments were drawn with replacement.

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 if 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 overtesting, 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. These indices reduce the total number of tests conducted. Finally, I adjust the *p*-values on the summary index tests to reflect the fact that I test multiple summary indices. Specifically, I control for Familywise Error Rate (FWE) using 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 outcomes tested 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). These tests feature two advantages over testing individual outcomes. First, they are robust to over-testing because the probability of Type I error does not increase as additional outcomes are added to a summary index. Second, they are potentially more powerful than individual level tests - multiple outcomes that approach marginal significance may aggregate into a single index that attains statistical significance. For example, consider an underlying latent index - human capital at a given age - that is expressed through multiple measures, such as years of education, employment, earnings, and criminal record. When testing whether preschool affects the latent index, two sources of random error exist. 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 into a single index.

At the most basic level, a summary index is a weighted mean of several standardized outcomes. The weights are calculated to maximize the amount of information captured in the index. To implement the summary index tests, I demean all outcomes and convert them to effect sizes by dividing each outcome by its control group 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 define three groupings, or "areas," of outcomes: pre-teen, adolescent, and adult. Every outcome y_{jk} is assigned to one of these three areas, resulting in a total of K_j outcomes in each area j.

I then create a new variable, $\overline{s}_{ij}$, that is the mean of the normalized, demeaned outcomes for child i in area j. When constructing $\overline{s}_{ij}$, the outcomes y_{ijk} are weighted by the inverse of the covariance matrix of the outcomes in area j (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 each other receive less weight; O'Brien (1984) finds this procedure to be more powerful than other popular tests in the repeated measures setting. Thus

$$\overline{s}_{ij} = \sum_{k \in \mathbb{K}_{ij}} w_{jk} \frac{y_{ijk} - \overline{y}_{jk}}{\sigma^{y}_{jk}}$$

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, σ^{y}_{jk} is the control group standard deviation, and w_{jk} is the outcome weight from the covariance matrix (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.

Each summary index consolidates several individual tests into a single test. However, there are still nine summary tests per gender. To address this problem, I 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. 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 summary index *p*-value, which occurs for teenage Perry females (Table 2). 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, 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, making the adjustment is less severe.

I implement the free step-down resampling method by first sorting the *M* tested outcomes (summary index tests in this case) in order of decreasing statistical significance. Let y_1,...,y_M be the sorted outcomes and p_1,...,p_M be their associated p-values. The sorting implies $p_1 < p_2 < ... < p_M$. I simulate the dataset under the null hypothesis of no treatment effect using the resampling procedure described earlier. I then calculate a set of simulated p-values, p_1^*,...,p_M^*, for the outcomes based on the simulated treatment status; these p-values will roughly follow the uniform distribution.

For each simulated p-value p_r^*, I construct a new p-value p_r^{**} such that $p_r^{**} = \min\{p_r^*, p_{r+1}^*, ..., p_M^*\}$ (note that *r* denotes the original significance rank of the outcome). Thus p_r^{**} follows the distribution of the minimum p-value from a set of $M - r + 1$ p-values (e.g., it is the minimum of *M* p-values for the most significant outcome, but the minimum of only one p-value for the least significant outcome). I repeat the simulation procedure 10,000 times. For each outcome y_r I tabulate S_r, the number of times that p_r^{**} is less than p_r. The FWE adjusted p-value for outcome y_r is then $\frac{S_r}{10{,}000}$ (subject to a final monotonicity enforcement that ensures that larger unadjusted p-values always correspond to larger adjusted p-values). The code for the procedure is available from the author.

3.2 Complications

Several complications, analyzed in-depth in Anderson (2006), 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, the direction of the induced bias is ambiguous, and standard corrections for missing data can be unreliable (cf. Paul, Mason, McCaffrey, and Fox, 2003). To address the attrition problem I therefore 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 affects females positively at the pre-teen stage. Table 2 reports summary index results by outcome stage and experiment. Like all tables in this section, it presents results for both genders. Coefficients in this table represent effect sizes; an effect size of 0.8 is generally considered large, 0.5 moderate, and 0.2 small (Cohen, 1988). 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. After adjusting for multiple inference, the Perry *p*-value remains significant, but the Abecedarian *p*-value falls just short of marginal significance. The Early Training females experience a summary effect size increase of 0.38, but the coefficient is insignificant. Males, however, do not experience consistent gains in pre-teen outcomes. Abecedarian males realize a summary effect size increase of 0.42, but it is insignificant after adjusting for multiple inference. The Perry and Early Training males experience summary effect size increases of 0.15 and 0.14 respectively; neither result approaches significance.

The disaggregated results suggest that preschool raises early IQ scores for both genders and reduces early grade retention and special education placement for females. However, preschool has limited effects on grade retention and special education for males.

Table 4 reports effects on pre-teen IQ scores. 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 (which in general are qualitatively similar to the standard 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 results in Table 5 suggest that the early IQ gains translate into better performance in primary school.[3] Female grade retention falls by 20 to 30 percentage points in all three programs, with *p*-values ranging from 0.08 to 0.16 (when interpreting the *p*-values for individual outcomes, note that they do not adjust for multiple inference). 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.

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.

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. Furthermore, programs may ”teach” towards specific tests (Klein, Hamilton, McCaffrey, and Stecher, 2000). Currie and Thomas (1995) and Garces, Thomas, and Currie (2002) conclude that, for African-Americans,

[3] 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.

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 therefore focus on long-term teenage and adult outcomes.

4.2 Teenage Outcomes

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 2). The Perry effect is highly significant (FWE-adjusted $p = 0.003$); the Abecedarian effect appears significant ($p = 0.04$), but it loses significance after adjusting for multiple inference. 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, and all are insignificant.

The disaggregated results suggest that early intervention improves high school graduation, employment, and juvenile arrest rates for females, but 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 Abecedarian, 49 percentage points in Perry, 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).

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 approximately 1,600 dollars less in annual government transfers at 19 ($p = 0.04$). 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.

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

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 2). 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 it is insignificant after adjusting for multiple inference. Early Training males experience a *decline* of 0.65 standard deviations in the adult summary index. This decrease appears significant ($p = 0.017$), but it is only marginally significant after FWE adjustments.

The disaggregated results suggest that preschool raises college attendance rates for females, improves female economic outcomes, and reduces female 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.

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). The negative effect for Early Training males is highly significant ($p = 0.005$), most likely due to over-testing.

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 ($p = 0.08$), though this effect disappears by age 40. Perry females earn more at ages 27 and 40 than their control counterparts (annual figures suggest approximately 2,600 and 3,500 dollars per year respectively, while monthly figures suggest about 400 and 160 dollars per month respectively), but the effects are mostly insignificant. Early Training 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 mixed 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 report insignificant increases in annual earnings of approximately 2,400 and 6,200 dollars at ages 27 and 40 respectively. In contrast, their reported monthly earnings increase by 537 dollars at age 27 ($p = 0.03$), but at age 40 the increase drops to 436 dollars and is insignificant. Perry males at age 40 experience a positive employment effect of 20 percentage points ($p = 0.11$). 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. 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 ($p = 0.07$).

There is some evidence that preschool affects marriage rates. At age 27, Perry females have a significantly higher marriage rate than untreated females. The 32 percentage point increase represents a 382 percent rise over the control group’s base rate ($p < 0.01$). Perry males, however, have the same marriage rate at 27 as their control counterparts.

Several female treatment effects are significantly higher than corresponding male effects, although the effect heterogeneity is less pronounced than during the teenage years. 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 partially the result of negative male treatment effects. Nevertheless, it still constitutes evidence of greater benefits for females - the female coefficients are centered around a higher mean, so even in the event of adverse shocks they do not become negative and significant.

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.[4] Treated females show particularly sharp increases in high school graduation and college attendance rates, but they also demonstrate 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. A few positive, long-term outcomes achieve or approach statistical significance for Perry males, including monthly earnings at age 27 and employment at age 40. However, these positive results are offset by several negative, significant outcomes for males, both in Perry and other programs.

A visual inspection of the results illustrates this pattern. Figure 1 presents a graphical 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 *t*-statistics generated by randomly assigning treatment

[4] 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.

status to females. 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. To construct this column, I randomly generate a set of treatment assignments and then compute and plot the corresponding *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 minimum value in the third column exceeds the minimum value in the first column, but the first column has more *t*-statistics clustered 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 inference.

A formal analysis examines summary index FWE *p*-values and aggregates all long-term outcomes into a single summary index. Females in the Abecedarian and Perry programs demonstrate significant improvements during the adult and teenage years respectively. In contrast, no male summary index achieves statistical significance (in the positive direction) after FWE adjustments.

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 female-male gap in treatment effects is consistent with previous findings in the non-experimental literature and reinforces a general perception that schooling helps girls more than boys (Tyre, 2006). 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 is beyond the scope of this paper, but 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.

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. One example is the role of teen pregnancy in high school dropouts. Since males cannot get pregnant, 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. For example, in developing countries it is common for families to invest more resources in boys than in girls (Bouis, et al., 1998). If a similar imbalance exists in the United States, and if preschool remedies this type of underinvestment, then girls might see greater gains from preschool than boys.

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 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 de novo analysis of the influential experimental preschool literature using statistical techniques that adjust for multiple inference. 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 in total years of schooling. The finding that preschool has a positive overall effect on females is significant in two of the three programs even after adjusting for multiple inference.

For males, however, there is no evidence of positive, long-term preschool treatment effects. Despite several positive and significant results, 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. Previous research has overlooked this finding because there has been no systematic analysis by gender across experiments and because no one has has applied a statistical framework that is robust to problems of multiple inference.

These results highlight both methodological and substantive points. First, they underscore the importance of multiple inference corrections in the context of the program evaluation literature. Many studies in this field test dozens of outcomes and focus on the subset of results that achieve significance. In response, the statistical framework presented in this paper enables researchers to address the issue of multiple testing while minimizing the loss in statistical power.

In addition, this paper makes clear several points in the context of the current human capital literature. 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.

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)

Anderson, Michael (2006) 'Uncovering Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects.' MIT Department of Economics, manuscript

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)

Bouis, Howarth, Marilou Palabrica-Costello, Orville Solon, Daniel Westbrook, and Azucena Limno (1998) *Gender Equality and Investments in Adolescents in the Rural Philippines* (International Food Policy Research Institute)

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

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)

Cohen, Jacob (1988) *Statistical Power Analysis for the Behavioral Sciences* (Lawrence Erlbaum Associates)

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

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

Efron, Bradley, and Robert Tibshirani (1994) *An Introduction to the Bootstrap* (CRC Press)

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

Fisher, R.A. (1935) *The Design of Experiments* (Hafner)

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

Hanushek, Eric (1986) 'The Economics of Schooling: Production and Efficiency in Public Schools.' *Journal of Economic Literature* 24(3), 1141–1177

— (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

Hochberg, Yosef (1988) 'A Sharper Bonferroni Procedure For Multiple Tests of Significance.' *Biometrika* 75(4), 800–802

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

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

Klein, Stephen, Laura Hamilton, Daniel McCaffrey, and Brian Stecher (2000) 'What Do Test Scores in Texas Tell Us?' RAND Issue Paper IP-202

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

Krueger, Alan (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)

Lewin, Tamar (2006) 'At Colleges, Women Are Leaving Men in the Dust.' *The New York Times* July 9, 2006, Section 1, 1

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)

Paul, Christopher, William Mason, Daniel McCaffrey, and Sarah Fox (2003) 'What Should We Do About Missing Data (A Case Study Using Logistic Regression with Missing Data on a Single Covariate).' California Center for Population Research Working Paper CCPR-028-03

Romano, Joseph, and Michael Wolf (2005) 'Stepwise Multiple Testing As Formalized Data Snooping.' *Econometrica* 73(4), 1237–1282

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)

Slavin, Robert (1989) 'Class Size and Student Achievement: Small Effects of Small Classes.' *Educational Psychologist* 24(1), 99–110

Stecher, Brian, Daniel McCaffrey, and Delia Bugliari (2003) 'The Relationship Between Exposure to Class Size Reduction and Student Achievement in California.' *Education Policy Analysis Archives* 11(40), 1–27

Tyre, Peg (2006) 'The Trouble With Boys.' *Newsweek* January 30, 2006, 44–52

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

Zellman, Gail, Brian Stecher, Stephen Klein, Daniel McCaffrey, Silvia Gutierrez, Rodger Madison, Denise Quigley, and Lisa Suarez (1998) *Findings From an Evaluation of the Parent Institute for Quality Education Parent Involvement Program* (RAND)

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: Summary Index Effects

		Female				**Male**				Gender
Project	Age	**Effect**	Naive	FWE	N	**Effect**	Naive	FWE	N	Interaction
			p-val	*p*-val			*p*-val	*p*-val		*t*-stat
ABC	Pre-Teen	0.445	0.026	0.117	54	0.417	0.026	0.187	51	0.11
		(0.194)				(0.181)				
Perry	Pre-Teen	0.537	0.004	0.026	51	0.150	0.387	0.941	72	1.53
		(0.177)				(0.172)				
ETP	Pre-Teen	0.380	0.170	0.346	30	0.142	0.557	0.960	34	0.67
		(0.270)				(0.238)				
ABC	Teen	0.422	0.042	0.153	53	0.162	0.407	0.941	51	0.93
		(0.202)				(0.194)				
Perry	Teen	0.613	0.000	0.003	51	0.035	0.716	0.976	72	3.32
		(0.156)				(0.096)				
ETP	Teen	0.551	0.104	0.346	29	0.097	0.781	0.976	32	0.95
		(0.327)				(0.345)				
ABC	Adult	0.452	0.003	0.022	53	0.312	0.066	0.369	51	0.64
		(0.144)				(0.166)				
Perry	Adult	0.358	0.022	0.117	51	-0.017	0.894	0.976	72	1.88
		(0.151)				(0.130)				
ETP	Adult	-0.067	0.723	0.709	29	-0.654	0.017	0.090	31	1.82
		(0.188)				(0.257)				

Notes: Parentheses contain OLS standard errors. Naive *p*-values are unadjusted *p*-values based on the *t*-distribution. FWE *p*-values adjust for multiple testing at the summary index level and are computed as described in Section (3). *t*-statistics test the difference between female and male treatment effects. See Table 3 for the components of each summary index.

Table 3: 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 4: 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 5: 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 (0.125)	0.429	0.082	53	-0.188 (0.142)	0.545	0.201	50	-0.21
Spec Educ	12	ABC	-0.066 (0.123)	0.296	0.576	53	-0.269 (0.140)	0.591	0.054	50	1.10
Repeat Grade	12	Perry	-0.201 (0.137)	0.409	0.134	46	0.078 (0.124)	0.389	0.514	66	-1.51
Spec Educ	17	Perry	-0.262 (0.129)	0.462	0.060	51	-0.037 (0.119)	0.462	0.741	72	-1.28
Retained	17	ETP	-0.284 (0.195)	0.600	0.156	29	0.100 (0.192)	0.600	0.514	30	-1.40
Special Help	17	ETP	0.116 (0.171)	0.200	0.529	29	0.036 (0.188)	0.364	0.832	31	0.31

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

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

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 (0.137)	0.536	0.432	53	0.188 (0.142)	0.455	0.196	50	-0.43
Employed	27	Perry	0.255 (0.136)	0.545	0.076	47	0.036 (0.121)	0.564	0.765	69	1.20
Annual Income	27	Perry	2,567 (2,686)	8,986	0.353	47	2,363 (2,708)	12,495	0.382	66	0.05
Monthly Income	27	Perry	396 (236)	651	0.102	47	537 (247)	830	0.027	68	-0.41
Employed	40	Perry	0.015 (0.115)	0.818	0.922	46	0.200 (0.120)	0.500	0.109	66	-1.12
Annual Income	40	Perry	3,492 (5,491)	17,374	0.536	46	6,228 (5,958)	21,119	0.302	66	-0.34
Monthly Income	40	Perry	162 (431)	1,615	0.715	46	436 (562)	1,839	0.459	66	-0.39
Receive Income	21	ETP	-0.074 (0.200)	0.600	0.688	29	-0.159 (0.134)	0.909	0.303	31	0.36
Receive Welfare	21	ETP	-0.042 (0.157)	0.200	0.805	30	N/A (N/A)	0.000	N/A	35	N/A

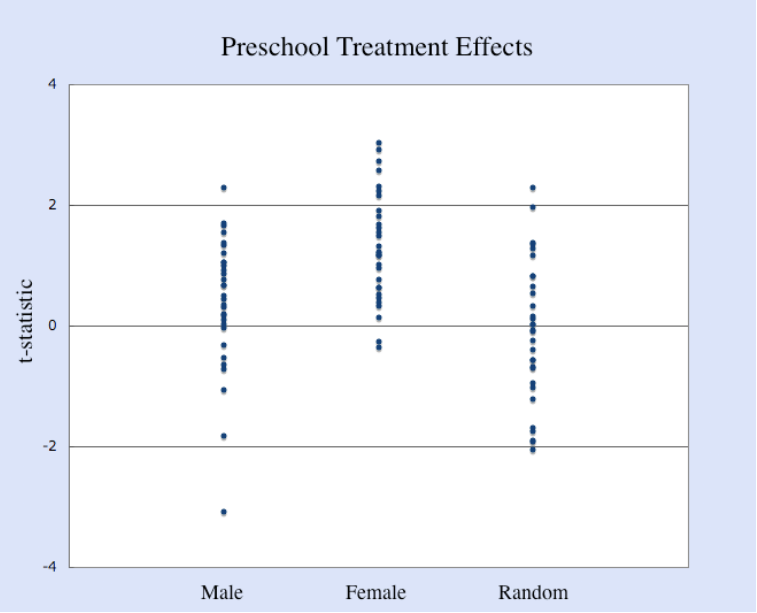
Notes: Parentheses contain robust standard errors. CM refers to control mean. Sample size varies within experiments due to attrition for some variables. *P*-values are computed as described in Section (3); *t*-statistics test the difference between female and male treatment effects.

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

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.

Figure 1: Effects of Preschool on Teen and Adult Outcomes


Preschool Treatment Effects
t-statistic
Male
Female
Random