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

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

Abstract

The view that the returns to educational investments are highest for early childhood interventions is widely held and stems primarily from several influential randomized trials – Abecedarian, Perry, and the Early Training Project – that point to super-normal returns to early interventions. This paper presents a de novo analysis of these experiments, focusing on two core issues that have received limited attention in previous analyses: treatment effect heterogeneity by gender and over-rejection of the null hypothesis due to multiple inference. To address the latter issue, I implement a statistical framework that combines summary index tests with Familywise Error Rate and False Discovery Rate corrections. The first technique reduces the number of tests conducted; the latter two adjust the p-values for multiple inference. The primary finding of the 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, which have appeared ambiguous when using "naive" estimators that fail to adjust for multiple testing, contribute to a growing literature on the emerging female-male academic achievement gap. They also demonstrate that in complex studies where multiple questions are asked of the same data set, it can be important to declare the family of tests under consideration and to either consolidate measures or report adjusted as well as unadjusted p-values.

Keywords: Program evaluation; Familywise error rate; Multiple comparisons; Preschool; False discovery rate

*Michael Anderson is Assistant Professor, Department of Agricultural and Resource Economics, University of California, Berkeley, CA 94720 (E-mail: mlanderson@berkeley.edu). Funding from the National Institute on Aging, through Grant Number T32-AG00186 to the NBER, is gratefully acknowledged. The author thanks Josh Angrist, David Autor, Jon Gruber, three anonymous referees, and an associate editor for their valuable insights, as well as Larry Schweinhart and Zongping Xiang of High/Scope, Frances Campbell and Elizabeth Pungello of UNC Chapel Hill, and Craig Ramey of Georgetown University for their generous assistance in obtaining the Perry Preschool Program and Abecedarian Project data used in this study. This research also used the Early Training Project, 1962-1979. These data were collected by Susan Walton Gray, and are available through the Henry A. Murray Research Archive at Harvard University, Cambridge, MA.

1 INTRODUCTION

The education literature contains dozens of papers showing inconsistent or low returns to publicly funded human capital investments (cf. Hanushek 1986; Stecher, McCaffrey, and Bugliari 2003). In contrast to these studies, several randomized early intervention experiments report striking increases in short-term IQ scores and long-term outcomes for treated children (Gray, Ramsey, and Klaus 1982; Campbell, Ramey, Pungello, Sparling, and Miller-Johnson 2002; Schweinhart, et al. 2005). These results have been highly influential and are often cited as proof of efficacy for many types of early interventions (cf. Currie 2001). 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 (Carneiro and Heckman 2003; Krueger 2003).

This paper focuses on the three prominent early intervention experiments: 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, and researchers conducted subject interviews and examined school and government records to collect long-term follow-up data on academic, social, and economic outcomes.

However, 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 received limited attention in the policy evaluation literature. 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 comprehensive statistical framework to directly address concerns about sample size and multiple inference. This general framework is broadly applicable to a range of program evaluation studies, which often have small samples and many outcomes. 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 early intervention programs separately by gender. The organization is as follows. Section 2 describes the data and each program's experimental design. Section 3 sets out the statistical framework. Section 4 presents results organized by outcome stage – preteen, teen, and adult – and benchmarks the performance of multiple inference adjustments when applied to a single study. Section 5 summarizes the main results and places them in the context of the broader literature. Section 6 concludes. The results demonstrate that early interventions (interventions that occur pre-kindergarten) significantly improve later-life outcomes for females, particularly academic achievement. However, treatment effects are modest or nonexistent for males – a fact that has been obscured when using "naive" analyses that fail to account for multiple inference.

2 EXPERIMENTAL BACKGROUND AND DATA

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 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 classes of about six. In contrast to the other programs, Abecedarian control children received some minor interventions: iron fortified formula, free diapers, and supportive social services when appropriate (Campbell and Ramey 1994). Of the three early intervention projects, Abecedarian was by far the most intensive.

The Abecedarian data set 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 12 and once at age 15. Researchers collected information on grade retention and special education at ages 12 and 15 from school records. Data on high school graduation, college attendance, employment, pregnancy, and criminal behavior come from an age 21 interview. Follow-up attrition rates are low, ranging from three to six percent for most outcomes.

2.2 The Perry Preschool Program

The Perry Preschool Program treated five waves of children in Ypsilanti, Michigan from 1962 to 1967. Children were randomly assigned to treated and control groups. Most treated children entered the program at age three and remained in it for two years; the first wave entered at age four and received one year of treatment. The program implemented the ideas of Jean Piaget and focused on language, socialization, numbers, space, and time in classes of five to six. Treated children attended the program five mornings per week from October through May and received one 90 minute home visit per week (Schweinhart, et al. 2005).

The Perry data set 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 10, and once more at age 14. 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 10 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 and participated in activities focusing on motivation and persistence in classes of four to five. They also received one 90 minute home visit per week for the program's duration.

The Early Training Project gathered data on 88 children. The study's control group consists of a local control group and a distal control group. Of the 88 children in the study, 61 lived in Murfreesboro, and 27 lived in another 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 sources: interviews with subjects and parents, program administered tests, and school records. IQ tests were given annually from ages four through eight and at ages 10 and 17. Data on grade retention and high school enrollment comes from school records. Subject interviews provide data on post-high school education and economic outcomes. No crime data were collected. Attrition rates for most variables are below 10 percent, and females 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 teen years range from 77.7 to 93.2. High school dropout rates range from 30 to 40 percent. In one sample, a majority of subjects have a criminal record. When drawing inferences about the results' external validity, it is important to note that these children are not representative of the average American child. Nevertheless, many of their attributes are not unusual for African-American youth in poor neighborhoods (cf. Miller 1992).

2.5 Internal Study Group Findings

Each study group has published manuscripts documenting the evolution of differences between treatment and control groups over time. In spite of substantial variation in treatment intensity across programs, similarities in outcome patterns emerge. All studies report significant, meaningful effects on IQ scores during the pre-kindergarten treatment period. These effects diminish over time, however, and by high school the IQ effects drop in magnitude by 70% to 100%. Nevertheless, all three studies report increases in schooling completion rates for treated children; high school graduation or college attendance rates rise by as much as 17 to 22 percentage points in each study. It therefore appears that although the cognitive benefits of these programs fade out, the non-cognitive benefits persist and manifest themselves in improved schooling completion rates later in life (Gray, et al. 1982; Schweinhart, et al. 1993; Campbell and Ramey 1994, 1995; Campbell, et al. 2002).

Nevertheless, important divergences appear between these studies' findings. In particular, the Perry Preschool Program reports large, statistically significant reductions in juvenile and adult criminal behavior that do not replicate in the Abecedarian Program. This divergence is not due to a low base rate of criminal behavior among the Abecedarian sample; the Abecedarian and Perry control groups display similar arrest rates (Schweinhart, et al. 1993; Clarke and Campbell 1998; Campbell, et al. 2002).

The findings become even more contradictory when effects are reported separately by

gender. The Early Training and Abecedarian programs do not consistently report effects by gender. For example, Gray, et al. (1982) report effects by gender for 5 of the 17 sets of results they present, while Campbell, et al. (2002) report treatment-by-gender interactions for 3 of the 15 adult demographic outcomes they present. Nevertheless, both study groups suggest in summary discussions that benefits for males may be modest. Early Training investigators caution that "as a whole, it looks as if the intervention program...was more effective for the females than the males" (Gray, et al. 1982, p. 254). Abecedarian researchers note that "treated women made greater educational progress relative to untreated women than was true for treated men relative to untreated men" and mention no significant long-term effects for males (Campbell, et al. 2002, p. 54).

The Perry Preschool manuscripts report effects separately by gender when results are significant. In contrast to the other studies, Perry investigators conclude there is no evidence of weaker benefits for males. In summarizing the overall benefits of the program, they state, "There is no suggestion that from a public policy perspective, preschool programs make sense for females but not for males, or vice versa" (Schweinhart, et al. 1993, p. 166). In fact, Schweinhart, et al. (2005) conclude that the total benefits for males are four times greater than the total benefits for females.

On the whole, there is therefore no consensus regarding the heterogeneity of early intervention effects by gender. This ambiguity may be due to the large numbers of outcomes tested in each study; every study group comes to a different conclusion because each one focuses on its subset of significant outcomes. In applying a framework that is robust to multiple inference, I untangle the conflicting gender-specific findings in the existing literature. Furthermore, I demonstrate that, when applied to a single study, these methods generate robust conclusions that replicate in the other two studies. This performance is encouraging and stands in contrast to the unstable conclusions produced by "naive" analyses.

3 STATISTICAL FRAMEWORK

3.1 Identification and Inference

The random assignment process makes estimation of causal effects straightforward. The primary approach compares treated children (those that received the intervention) to untreated children (those that did not) across a wide variety of outcomes. 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 10 individuals. The Huber-White standard errors may therefore be misleading, particularly since the underlying data is distributed non-normally in some cases. 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 (cf. Efron and Tibshirani 1993). This procedure computes the null distribution of the test statistic under minimal assumptions: random assignment and no treatment effect. For a given sample size N_k, the procedure is implemented as follows:

1. Draw binary treatment assignments z_i^* from the empirical distribution of the original treatment assignments without replacement.

2. Calculate the *t*-statistic for the difference in means between treated and untreated groups.

3. Repeat the procedure 100,000 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, reject the null hypothesis of no treatment effect. This procedure tests the sharp null hypothesis of no treatment effect, so rejection implies that the treatment has some distributional effect. Formally, the two required assumptions are:

1. Random Assignment: Let y_{i0} be the outcome for individual i when untreated and y_{i1} be the outcome for individual i when treated (we only observe either y_{i0} or y_{i1}). Random assignment implies $\{y_{i0}, y_{i1} \perp z_i\}$.

2. No Treatment Effect: $y_{i0} = y_{i1} \; \forall \; i$

Note that no assumptions regarding the distributions or independence of potential outcomes are needed. This is because the randomized design itself is the basis for inference (Fisher 1935), and pre-existing clusters cannot be positively correlated with the treatment assignments in any systematic way. Even if the potential outcomes are fixed, the test statistic will still have a null distribution induced by the random assignment. Since the researcher knows the design of the assignment, it is always possible to reconstruct this distribution under the null hypothesis of no treatment effect, at least by simulation if not analytically. Thus, this test always controls Type I error at the desired level (Rosenbaum 2007).

For binary y_i, this test generally converges to Fisher's Exact Test. However, it differs slightly from Fisher's Exact Test in that Fisher's test rejects for small *p*-values while this test rejects for large *t*-statistics. This test is also similar to bootstrapping under the assumption of no treatment effect (Simon 1997); the only difference is that the resampling is done without replacement rather than with replacement. This highlights the fact that the variance in the test statistic's null distribution arises from the randomization procedure itself rather than from unknown variability in the potential outcomes.

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.

3.2 Multiple Inference Adjustments

Several papers in the educational field have discussed the issue of simultaneous inference with large numbers of outcomes (cf. Williams, Jones, and Tukey 1999), and some research organizations, such as the Institute of Education Sciences' What Works Clearinghouse,

have technical standards that include multiplicity adjustments. However, most randomized evaluations in the social sciences test many outcomes but fail to apply any type of multiple inference correction. To gauge the extent of the problem, I conducted a survey of randomized evaluation papers published from 2004 to 2006 in the fields of economic or employment policy, education, criminology, political science or public opinion, and child or adolescent welfare. Using the *CSA Illumina* social sciences databases, I identified 44 such papers in peer-reviewed journals.

Of these 44 articles, 37 (84%) report testing five or more outcomes, and 27 (61%) report testing ten or more outcomes. These figures represent lower-bounds for the total number of tests conducted, since many tests may be conducted but not reported. Nevertheless, only three papers (7%) implement any type of multiple inference correction. Of these three papers, two apply the Bonferroni correction – the most rudimentary adjustment in general use – and one implements a summary index that reduces the total number of tests. Although multiple inference corrections are standard (and often mandatory) in psychological research (Benjamini and Yekutieli 2001), they remain uncommon in other social sciences, perhaps because practitioners in these fields are unfamiliar with the techniques or because they have seen no evidence that they yield more robust conclusions.

Two approaches exist to solving the multiple inference problem. One approach reduces the number of tests being conducted. This method avoids p-value adjustments, which generally reduce the power of any given test, at the cost of limiting the scope of hypothesis testing. The other approach maintains the number of tests but adjusts the p-values to reflect this fact. This method allows for an arbitrarily large number of tests, but the power of each specific test can fall as the number of tests conducted grows. In this paper, I combine both approaches in order to balance the trade-offs of each one.

I begin by limiting the total number of hypotheses being tested. First, I choose a specific set of outcomes based on a priori notions of importance. I then implement summary index tests in three broad outcome areas: preteen, adolescent, and adult. These indices combine multiple measures to reduce the total number of tests conducted.

Nevertheless, I still test multiple indices. I therefore adjust the p-values on the summary

index tests to reflect this fact. Specifically, I control *Familywise Error Rate* (FWER) – the probability of rejecting at least one true null hypothesis – using the free step-down resampling method. When reporting results for specific outcomes, I control the *False Discovery Rate* (FDR), or the proportion of rejections that are "false discoveries" (Type I errors). FDR control is well suited to exploratory analysis because it allows a small number of Type I errors in exchange for greater power than FWER control.

3.2.1 Summary Index Tests

In this study I define a set of primary outcomes that includes IQ scores, grade retention, special education, high school graduation, college attendance, employment, earnings, government transfers, arrests, convictions or incarcerations, drug use, teen pregnancy, and marriage (see Table 2). This list appears long but represents only a small fraction of all available outcomes. Nevertheless, the total number of outcomes tested reaches 47. I therefore implement summary index tests that pool multiple outcomes into a single test.

Summary index tests originate in the biostatistics literature (see O'Brien 1984). These tests feature three advantages over testing individual outcomes. First, they are robust to over-testing because each index represents a single test. Therefore, the probability of a false rejection does not increase as additional outcomes are added to a summary index. Second, they provide a statistical test for whether a program has a "general effect" on a set of outcomes. Finally, 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 variable – 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 early intervention affects the latent variable, two sources of random error exist. First, there is error that arises from the random assignment procedure – the latent variable 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 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. A summary index test can be implemented through the following steps (see Appendix A for a formal definition):

1. For all outcomes, switch signs where necessary so that the positive direction always indicates a "better" outcome.

2. Demean all outcomes and convert them to effect sizes by dividing each outcome by its control group standard deviation. Call the transformed outcomes $\tilde{y}$. (This conversion normalizes outcomes to be on a comparable scale.)

3. Define J groupings of outcomes (also referred to as areas or domains). Every outcome y_{jk} is assigned to one of these J areas, giving K_j outcomes in each area j (k indexes outcomes with an area).

4. Create a new variable, $\overline{s}_{ij}$, that is a weighted average of $\tilde{y}_{ijk}$ for individual i in area j. When constructing $\overline{s}_{ij}$, weight its inputs, outcomes $\tilde{y}_{ijk}$, by the inverse of the covariance matrix of the transformed outcomes in area j. (A simple way to do this is to set the weight on each outcome equal to the sum of its row entries in the inverted covariance matrix for area j. Formally, $\overline{s}_{ij} = (\mathbf{1}'\hat{\mathbf{\Sigma}}_j^{-\mathbf{1}}\mathbf{1})^{-1}(\mathbf{1}'\hat{\mathbf{\Sigma}}_j^{-\mathbf{1}}\tilde{\mathbf{y}}_{\mathbf{ij}})$, where $\mathbf{1}$ is a column vector of ones, $\hat{\mathbf{\Sigma}}_j^{-\mathbf{1}}$ is the inverted covariance matrix, and $\tilde{\mathbf{y}}_{\mathbf{ij}}$ is a column vector of all outcomes for individual i in area j. Note that this is an efficient generalized least squares (GLS) estimator.)

5. Regress the new variable, $\overline{s}_{ij}$, on treatment status to estimate the effect of treatment on area j. A standard t-test assesses the significance of the coefficient.

In this research I define three groupings based on age: preteen, adolescent, and adult. Given the interest in these programs' long-term impacts, testing for effects at the adolescent and adult stages is natural. Nevertheless, the choice of outcome groupings can theoretically affect the results, so one should check that results are robust to alternative grouping choices.

For example, in this paper grouping outcomes by academic, economic, and social domains, rather than stage-of-life domains, does not qualitatively change the results. (If the results are sensitive to grouping choice, then summary index *p*-values should be adjusted using the techniques in Section 3.2.2 or 3.2.3 to reflect the fact that the most significant specification was chosen.)

The GLS weighting procedure in step 4 increases efficiency by ensuring that outcomes which are highly correlated with each other receive less weight, while outcomes that are uncorrelated and therefore represent new information receive more weight. O'Brien (1984) finds this procedure to be more powerful than other popular tests in the repeated measures setting. Also, 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.

3.2.2 Familywise Error Rate Control

Each summary index consolidates several individual tests into a single test. However, we may wish to test for effects in several domains or across multiple experiments, resulting in multiple summary indices. In this research, there are nine summary indices per gender (three domains by three experiments). One option is to further reduce the number of tests by aggregating all summary indices together. However, because differential effects by domain may be of interest, there is substantial benefit to maintaining separation between the indices. For example, long-term outcomes may be of greater policy interest than short-term test score gains. Therefore, an alternative approach is to maintain the number of summary indices and adjust their *p*-values to reflect the multiple inference problem.

The most common approach to adjusting *p*-values for multiple testing is to control Familywise Error Rate. Suppose a family of *M* hypotheses, $H_1, H_2, ..., H_M$, is tested, of which J are true ($J \leq M$). FWER is the probability that at least one of the *J* true hypotheses in the family is rejected. In this research, the family of tested hypotheses is the set of nine summary index tests performed for each gender. As more hypotheses are added to a family, the probability of rejecting at least one of them at a given α-level increases, and hence FWER increases. FWER control techniques adjust the *p*-values of each test upwards

to reduce the probability of a false rejection.

A popular technique for controlling FWER is the Bonferroni correction. This technique multiplies each *p*-value by M, the number of tests performed. Its advantage is simplicity, but it suffers from poor power. A more powerful technique that controls FWER is the free step-down resampling method (Westfall and Young 1993). This algorithm is more powerful than the Bonferroni correction (and other algorithms) for three reasons. First, the free step-down resampling method computes an exact probability rather than an upper bound (it is common, for example, for Bonferroni *p*-values to exceed 1). Second, when a hypothesis is rejected, the free step-down resampling method removes it from the family being tested, increasing the power of the remaining tests. Bonferroni does not. Finally, unlike Bonferroni, free step-down resampling incorporates dependence between outcomes. This can substantially increase power if outcomes are highly correlated. In an extreme case, if all outcomes are perfectly correlated, FWER adjusted p-values and the unadjusted p-values should be equal, and with the free step-down resampling method they will be.

For a family of M outcomes tested in an experimental setting, the free step-down resampling procedure is implemented as follows:

1. Sort outcomes $y_1, ..., y_M$ in order of decreasing significance (increasing *p*-value), i.e. such that $p_1 < p_2 < ... < p_M$.

2. Simulate the data set under the null hypothesis of no treatment effect using the resampling procedure described in Section 3.1.

3. Calculate a set of simulated *p*-values, $p_1^*, ..., p_M^*$, for outcomes $y_1, ..., y_M$ using the simulated treatment status variable. Note that they will not display the same montonicity as $p_1, ..., p_M$.

4. Enforce the original monotonicity: Compute $p_r^{**} = \min\{p_r^*, p_{r+1}^*, ..., p_M^*\}$. (*r* denotes the original significance rank of the outcome, with $r = 1$ being the most significant and $r = M$ being the least significant)

5. Perform $L \geq 100,000$ replications of steps 2 through 4. For each outcome y_r, tabulate S_r, the number of times that $p_r^{**} < p_r$.

6. Compute $p_r^{fwer*} = S_r/L$.

7. Enforce monotonicity a final time: $p_r^{fwer} = \min\{p_r^{fwer*}, p_{r+1}^{fwer*}, ..., p_M^{fwer*}\}$. (This final monotonicity enforcement ensures that larger unadjusted *p*-values always correspond to larger adjusted *p*-values.)

The crucial steps of this algorithm are steps 2 through 4. Steps 2 and 3 ensure that the dependence structure between outcomes is preserved because each case is resampled with the correlation structure of its outcomes intact. We therefore expect $p_1^*, ..., p_M^*$ to be positively correlated (if the original outcomes were positively correlated), and the minimum *p*-value of a set of M positively correlated *p*-values is generally greater than the minimum *p*-value of a set of M independent *p*-values. Incorporating dependence thus increases the probability that $p_r < p_r^{**}$, reducing S_r and increasing the probability of rejection.

Step 4 performs the key multiplicity adjustment when the simulated *p*-value for outcome y_r p_r^*, is replaced with $\min\{p_r^*, p_{r+1}^*, ..., p_M^*\}$. The original *p*-value, p_r, is thus judged against the distribution of the minimum *p*-value of a set of $M - r + 1$ *p*-values. This makes the adjusted *p*-value more conservative than a standard *p*-value, which is implicitly judged against the distribution of the minimum *p*-value of a set of one *p*-value, but less conservative than the Bonferroni correction, which implicitly judges every *p*-value against the distribution of the minimum *p*-value of a set of M *p*-values.

An example may aid interpretation of FWER adjusted *p*-values. In this research, there are $M = 9$ summary indices tested for each gender. Consider the smallest summary index *p*-value of the nine male summary indices, which occurs for adult Early Training males (Table 3). The unadjusted *p*-value is approximately 0.011. The corresponding adjusted *p*-value, calculated via the free step-down resampling method for the entire family of male summary tests, is $p^{fwer} = 0.090$. Suppose we simulate the male data 100,000 times under the null hypothesis of no treatment effect. If we compute an entire set of nine 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 0.011 approximately 9 percent of the time. A minimum observed p-value of 0.011 is therefore not unlikely under the null given the number of tests conducted – a fact that helps explain why this particular effect goes in the "wrong" (negative) direction. For unadjusted p-values above the family's minimum p-value, the number of tests in the family effectively decreases, making the adjustment less severe.

The free step-down resampling method strongly controls FWER – for any subset of the family of hypotheses, it ensures that the probability of falsely rejecting at least one hypothesis is less than α even if some of hypotheses outside of that subset are false (weak control of FWER only guarantees the size of a test if every hypothesis in the family is true). The only assumption necessary for this algorithm to provide strong control is subset pivotality, or the assumption that the distribution of any subset of the family of test statistics depends only on the validity of the hypotheses in that subset. For tests of multiple outcomes, such as this one, that assumption is met (Westfall, Tobias, Rom, Wolfinger, and Hochberg 1999, p. 237).

3.2.3 False Discovery Rate Control

FWER control limits the probability of making *any* Type I error. It is thus well suited to cases in which the cost of a false rejection is high. In this research, for instance, incorrectly concluding that early interventions are effective could result in a large-scale misallocation of teaching resources. However, in exploratory analysis we may be willing to tolerate some Type I errors in exchange for greater power. For example, the effects of early intervention on specific outcomes may be of interest, and since overall conclusions about program efficacy will not be based on a single outcome, it seems reasonable to accept a few Type I errors in exchange for greater power. This tradeoff is particularly appealing when, as in this case, we are testing a large number of hypotheses, because FWER adjustments become increasingly severe as the number of tests grows – it is inherent in controlling the probability of making a single false rejection. An alternative method of addressing the multiplicity problem that often affords better power is to control the False Discovery Rate, or the expected proportion of rejections that are Type I errors. FDR formalizes the tradeoff between

correct and false rejections and reduces the penalty to testing additional hypotheses.

Define V as the number of false rejections, U as the number of correct rejections, and $t = V + U$ as the total number of rejections. FWER is the probability that V is greater than 0. FDR is the expected proportion of all rejections that are Type I errors, or $E[Q = V/t]$ (when $t = 0$, Q is defined to be 0). If all null hypotheses are true, then $V = t$, and FWER and FDR are equivalent (Q equals 0 when there are no rejections and 1 when there are one or more rejections, so FDR = $E[Q] = P(t > 0) = P(V > 0)$ = FWER). However, when some false hypotheses are correctly rejected, then FDR is less than FWER because the expected proportion of rejections that are Type I errors is less than the probability of making any Type I error. Controlling FDR at a given level therefore often requires less stringent *p*-value adjustments than controlling FWER at the same level, resulting in increased power.

Benjamini and Hochberg (1995) propose a simple method for controlling FDR (referred to as BH from this point on). As in Section 3.2.2, suppose that we test hypotheses $H_1, ..., H_M$, and let the hypotheses be sorted in order of decreasing significance, such that $p_1 < p_2 < ... < p_M$. Suppose $q \in (0, 1)$. Let c be the largest r for which $p_r < qr/M$. Rejecting all hypotheses $H_1, ..., H_c$ controls FDR at level q for independent or positively dependent *p*-values. (In other words, beginning with p_M, check whether each *p*-value meets $p_r < qr/M$. When one does, reject it and all smaller *p*-values.) This procedure is in fact conservative in that it controls FDR at level $q(m_0/M)$, where m_0 is the number of true null hypotheses (Benjamini and Yekutieli 2001). We do not observe m_0, but if we did we could "sharpen" the procedure by replacing qr/M with qr/m_0. Since $qr/m_0 \geq qr/M$, the sharpened procedure would provide greater power if at least one null hypothesis were false.

Benjamini, Krieger, and Yekutieli (2006) propose a two-stage procedure that estimates the number of true hypotheses to achieve sharpened FDR control. The procedure is implemented as follows:

1. Apply the BH procedure at level $q' = q/(1 + q)$. Let c be the number of hypotheses rejected. If $c = 0$, stop. Otherwise, continue to step 2.

2. Let $\hat{m}_0 = M - c$.

3. Apply the BH procedure at level $q^* = q'M/\hat{m}_0$.

By incorporating the number of hypotheses rejected in the first stage into the second stage, this procedure provides better power than the standard BH procedure while controlling FDR at level q for independent *p*-values. Simulations indicate that the two-stage procedure also works well for positively dependent *p*-values (Benjamini, et al. 2006), such as the ones in this research. I therefore use the two-stage procedure to control FDR when reporting results for specific outcomes (e.g., high school graduation, employment, etc.). However, researchers dealing with negatively dependent *p*-values may need to adopt a more conservative modification of the BH procedure (Benjamini and Yekutieli 2001, p. 1169).

The BH and two-stage procedures both report whether a hypothesis was rejected at level q, but they do not report the smallest level q at which the hypothesis would be rejected. This value – which is the natural analog to the standard *p*-value – can easily be computed for all hypotheses by performing the procedure for all possible q levels (e.g., 1.000, 0.999, 0.998,...) and recording when each hypothesis ceases to be rejected. Stata code is available from the author to calculate these FDR "*q*-values."

To understand in practice why FDR control is less conservative than FWER control, consider how the BH and free step-down resampling procedures treat the median *p*-value, $p' = p_{M/2}$, in a set of M *p*-values. Roughly, the BH procedure rejects $H' = H_{M/2}$ if $p_{M/2} < \alpha(M/2)/M = \alpha/2$, while the free step-down resampling procedure rejects $H_{M/2}$ if $p_{M/2}$ exceeds the minimum of a family of $M/2$ simulated *p*-values at a rate less than α. The former equates to adjusting the *p*-value by a factor of 2, while the latter equates to adjusting the *p*-value by a factor of up to $M/2$. For large M, the difference becomes substantial. Note also that M does not appear on the right side of the expression $p_{M/2} < \alpha/2$. If additional *p*-values – distributed similarly to the existing *p*-values – are added to the family of tests, the FDR adjustment to the existing *p*-values need not become more stringent in expectation.

3.2.4 Summary

Three types of multiple inference adjustments are presented (and applied): summary index tests, FWER adjusted *p*-values, and FDR adjusted *p*-values. The first technique reduces

the total number of tests performed, while the second and third techniques maintain the number of tests and adjust the p-values. Given the substantial differences between these techniques, it is important that researchers understand the benefits and drawbacks of each technique when deciding which ones are most appropriate for their own work.

Summary index tests make sense when testing for an intervention's overall effect and when there is an a priori reason to believe that a group of outcomes will be affected in a consistent direction. In those cases, a summary index test often has better power than a series of FWER or FDR adjusted individual tests. This research applies summary indices to estimate the overall effects of each program at different stages in life.

Athough they are more likely to reject, summary index tests yield less information when they do reject, as it is impossible to conclude which underlying outcomes were significantly affected. If effects on specific outcomes are of interest, or if there is no reason to believe that outcomes are affected in a consistent direction, then testing all outcomes of interest and adjusting the p-values is a logical strategy. In that case, the choice between FWER and FDR adjustments may be dominated by the cost of a Type I error. When controlling FDR with many outcomes, one can expect to encounter some false positives with reasonably high probability. In contrast, when controlling FWER, all rejections will be correct with high probability. Therefore, if the cost of a Type I error is high, a researcher will likely opt for FWER control. However, if the cost of a Type I error is low to moderate, the increased power of FDR control will be appealing, particularly if the family of hypotheses being tested is large. This research applies FWER adjustments to the summary index p-values to ensure that programs are not erroneously judged to be effective at different life stages. It applies FDR adjustments to tests of individual outcomes to facilitate exploratory analysis while controlling the number of false rejections. Conclusions about overall program effectiveness, however, should be based upon the FWER adjusted summary index p-values.

4 RESULTS

4.1 Graphical Analysis

Figure 1 presents a graphical summary of the treatment effect t-statistics for long-term outcomes. This figure plots t-statistics for teenage and adult coefficients across all experiments for each gender (see rows marked "Teen" and "Adult" in Table 2). 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, suggesting females accrue greater long-term benefits from these programs.

The third column of points plots a set of t-statistics generated by randomly assigning treatment status to children and computing the corresponding t-statistics. This procedure guarantees that any significant "treatment effects" visible in the column are simply due to chance. The procedure is equivalent to sampling randomly from the t-distribution, except that it preserves the inherent correlation between t-statistics within each experiment.

The second and third columns are immediately distinguishable from each other, implying that females realize long-term benefits from these programs. Comparing the first and third columns, however, reveals that the distribution of male t-statistics is hard 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 both the first and third columns a case could be made for positive treatment effects by focusing on the set of outcomes near the top. This fact highlights the importance of correcting for multiple inference.

The following subsections analyze program effects by life-stage and experiment, as well as exploring effects for specific outcomes. I define two families of tests for calculating FWER and FDR adjusted p-values – one for each gender. (All female outcomes constitute one family, and all male outcomes constitute a second family. A case can be made for analyzing Abecedarian – the most intensive program – as a separate family; however, doing

so does not change the paper's central conclusions.) The reported summary effects control for FWER, or the probability of any false rejection, while the effects for specific outcomes control for FDR, or the expected proportion of false discoveries.

4.2 Preteen Outcomes

The interventions affect females positively at the preteen stage. Table 3 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. For comparison, the average effect size of a wide range of elementary school interventions summarized in Hill, Bloom, Black, and Lipsey (2007) is 0.33, and the black-white test score gap corresponds to an effect size of 0.8 to 1.0. At the preteen stage, the programs improve outcomes for Abecedarian and Perry females, with summary effect size increases of 0.45 and 0.54 respectively. Controlling FWER using the free step-down resampling method, the Perry p-value is significant, but the Abecedarian p-value falls short of marginal significance. Early Training females experience an insignificant summary effect size increase of 0.36.

Males, however, do not experience consistent gains in preteen outcomes. Abecedarian males realize a summary effect size increase of 0.42, but it is insignificant when adjusting for multiple inference. The Perry and Early Training males experience summary effect size increases of 0.15; neither result approaches significance.

The disaggregated results suggest that the interventions raise early IQ scores for both genders and reduce early grade retention and special education for females. However, they have limited effects on grade retention and special education for males.

Table 4 reports effects on preteen 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), the fourth column reports FDR "q-values" (computed using the two-stage procedure from Section 3.2.3), and the fifth column reports sample size. The last column in each table tests for differences between female and male treatment effects.

All projects demonstrate similar IQ effects at early ages. In each project, there is a large

IQ effect for at least one gender upon completion of preschool; in five cases – including two cases for males – results are significant when controlling FDR at $q = 0.10$. Females continue to display large IQ effects at age 10 in Abecedarian and Early Training. Males, however, display no significant IQ effect in any project at age 10.

The results in Table 5 suggest that the early IQ gains may translate into better performance in primary school, but no result rejects when controlling FDR at $q = 0.10$. Female grade retention falls by 20 to 30 percentage points in all three programs, and female special education placement falls 26 percentage points in the Perry program. Abecedarian males experience (insignificant) 19 and 27 percentage point declines in grade retention and special education placement. However, males in the Perry and Early Training programs demonstrate no notable decreases in grade retention or special education placement.

Gender differences in treatment effects emerge by age 10. Female IQ effects at age 10 are higher than 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. Most importantly, in every experiment the summary female preteen effect is higher than the summary male preteen effect.

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

4.3 Teenage Outcomes

Overall, the interventions have consistent, positive effects on female teen outcomes. Teen summary effects increase by 0.42, 0.61, and 0.46 standard deviations for females in the Abecedarian, Perry, and Early Training programs (see Table 3). The Perry effect is highly

significant ($p < 0.001$, $p^{fwer} = 0.003$). The interventions, however, have no significant effect on male teen outcomes; male summary effects increase by only 0.16, 0.04, and 0.12 respectively in the Abecedarian, Perry, and Early Training programs.

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 teen academic outcomes, including IQ scores and high school graduation rates. By age 14, initial IQ effects dissipate in all three programs. However, the minimal IQ effects belie strong gains among females for several important teen 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, 49, and 29 percentage points in Abecedarian, Perry, and Early Training respectively. The Perry result is highly significant ($p < 0.001$, $q = 0.001$). The Abecedarian and Early Training results, however, do not reject when controlling FDR at $q = 0.10$.

Male high school graduation effects, however, 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. No effect is significant.

Table 7 presents results for teenage economic and social outcomes. Females appear to experience positive economic effects from at least one intervention as teenagers. In Perry, treated females have teen unemployment rates that are 31 percentage points lower than untreated females ($p = 0.03$, $q = 0.11$). Treated females also receive roughly 1,600 dollars less in annual government transfers at 19 ($p = 0.04$, $q = 0.13$). Males, however, derive no significant economic benefits from the interventions during their teenage years.

One program has a significant effect on female teen criminal behavior; Perry females are 34 percentage points less likely to have a juvenile record ($p = 0.01$, $q = 0.05$). However, this result is not mirrored among males.

During the teenage years, it is clear that females benefit more than males from these interventions. The female-male difference in high school graduation effects is substantial in the Abecedarian and Perry programs ($t = 3.32$). Female-male differences also emerge

among Perry teens for effects on unemployment, criminal behavior, and government transfers. 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. With the exception of Abecedarian IQ scores, every reported teen effect is greater for females than for males.

4.4 Adult Outcomes

Overall, females benefit from at least one of the programs as adults. In the Abecedarian and Perry programs, females display positive general effects of 0.45 and 0.35 standard deviations respectively (see Table 3); the former effect is statistically significant ($p < 0.01$, $p^{fwer} = 0.02$). However, Early Training females demonstrate no general treatment effect as adults. This could be due to differences in the Early Training Project's intervention program, or it could be due to low statistical power.

Unlike females, males show little evidence of positive effects as adults. Summary effects for Abecedarian and Perry males increase by 0.31 and -0.01 standard deviations. The Abecedarian result appears marginally significant ($p = 0.07$), but in fact is insignificant ($p^{fwer} = 0.37$). Early Training males experience a *decline* of 0.71 standard deviations in the summary index. This decrease – due primarily to low college attendance rates among Early Training Males – appears highly significant ($p = 0.01$), but in fact is only marginally significant ($p^{fwer} = 0.09$). This unexpected finding in the "wrong" direction underscores the importance of multiplicity adjustments.

The disaggregated results suggest that, for females, early intervention may raise college attendance rates, improve economic outcomes, and reduce criminal behavior. The effects for males, however, are weaker 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. No male effect is statistically significant at levels of 0.05 or less after FDR adjustment. The discussion therefore focuses on possible female effects.

Table 8 reports treatment effects on college attendance. Early intervention may increase

the probability of college attendance for females. Abecedarian females report college attendance rates that are 29 percentage points higher than their control counterparts ($p = 0.02$, $q = 0.08$). Perry and Early Training post-high school education attendance rates increase by 12 to 16 percentage points, though neither effect is significant.

Table 9 reports results for adult economic outcomes. There is weak evidence of a positive effect on female economic outcomes. Perry females are 26 percentage points more likely to be employed at age 27 ($p = 0.08$, $q = 0.22$), and they earn more at ages 27 and 40 than their control counterparts (though these effects are statistically insignificant). Early Training females are less likely to receive welfare at age 21, but the effect is insignificant. It is possible that potential employment effects at age 21 for Abecedarian and Early Training women are masked by increased college attendance rates. However, controlling for college attendance does not appreciably change the employment coefficients for either program.

Table 10 presents effects on adult social behavior. Treated females report some reductions in criminal behavior. Abecedarian females are 32 percentage points less likely to use marijuana ($p < 0.01$, $q = 0.05$), though they experience no significant reduction in conviction or incarceration rates by age 21. Perry females have 86 percent fewer lifetime arrests (-1.95 arrests per capita, $p = 0.01$, $q = 0.07$), though they are only 15 percentage points less likely to have a criminal record.

There is some evidence that early intervention 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 ($p = 0.01$, $q = 0.07$).

Female treatment effects are generally higher than corresponding male effects, although the effect heterogeneity is less pronounced than during the teen years. The difference in female-male summary effects is substantial in Perry and the Early Training Project. Large female-male treatment effect differences emerge for drug use and marriage among Perry participants and post-high school education among Early Training participants. For drug use and post-high school education, the differential is partially due to negative male treatment effects. Nevertheless, it still constitutes evidence of greater benefits for females – the

female effects are centered around a higher mean, so even in the event of adverse shocks they do not become negative and significant.

4.5 Perry Reanalysis

As a final demonstration of the value of correcting for multiple inference, I conduct a stand-alone reanalysis of the Perry Preschool Project – arguably the most influential of the three experiments. For both male and female effects, I use the point estimates and standard errors for all Perry outcomes presented in Tables 4 through 10. I compute FDR q-values (not shown in tables) using all Perry outcomes as the family of tests under consideration, as the original Perry researchers would have done were they to apply this technique.

Under these conditions, I find that two effects – early male IQ scores and female high school graduation rates – reject when controlling FDR at $q = 0.05$. Three more effects – early female IQ scores, female marital rates, and female juvenile arrest rates – reject when controlling FDR at $q = 0.10$. Do these findings replicate in the other two studies? In general, yes. The early male IQ effect replicates strongly in Abecedarian. The female high school graduation effect replicates in both Abecedarian and Early Training, and the early female IQ effect replicates weakly in Abecedarian and strongly in Early Training. The only conclusion that fails to replicate is the female juvenile arrest rate effect, with a FDR q-value of 0.07. (No data on adult marital rates are available for Abecedarian and Early Training). A simple application of the two-stage FDR procedure – which requires no resampling and can even be implemented in a spreadsheet – therefore proves sufficient to generate robust conclusions that replicate in independent studies.

Now consider a conventional research design based on unadjusted p-values. Rejecting effects with "naive" (unadjusted) p-values of less than 0.10 adds eight more significant or marginally significant outcomes – female adult arrests, female employment, male monthly income, female government transfers, female special education rates, male drug use (in the adverse direction), male employment, and female monthly income. Of these eight outcomes, two (male and female monthly income) are not included in the other two studies. The remaining six fail to replicate in either of the other studies. The sharp contrast in

replication performance between findings that reject when controlling FDR and findings that reject based on unadjusted p-values emphasizes the benefits of applying even simple adjustments for multiple inference.

5 DISCUSSION

A clear pattern emerges from a detailed examination of treatment effects by gender: females display significant long-term effects from the interventions, while males show weaker and inconsistent effects. Treated females show particularly sharp increases in high school graduation and college attendance rates, but there is also evidence of positive effects for economic outcomes, criminal behavior, drug use, and marriage.

In contrast to females, males appear not to derive lasting benefits from the interventions. A few positive, long-term outcomes achieve or approach significance for Perry males (when using unadjusted p-values), including monthly earnings at 27 and employment at 40. However, these positive results are offset by several negative, significant male outcomes, both in Perry and other programs.

A summary test that pools all teen outcomes together across experiments finds an overall effect size of 0.51 for females (standard error of 0.13) and 0.08 for males (standard error of 0.14). The gender difference is significant ($p = 0.029, p^{fwer} = 0.029$). 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 significant ($p = 0.027, p^{fwer} = 0.029$). (FWER p-values are adjusted for the fact that gender differences are tested as teens and adults.) Of course, we can never reject arbitrarily small effects for males, and precision is limited by the relatively small samples. Some point estimates are of notable magnitude despite being insignificant. It is also of note that summary effects for males are larger at every stage in Abecedarian than Perry and Early Training. Perhaps males retain some benefits from highly intensive programs. Nevertheless, the overall results indicate that positive male treatment effects are likely modest at best.

The results help clarify several inconsistencies in the previous literature. First, they establish that girls benefited more than boys from these interventions – previous findings demonstrating significant long-term effects for boys, primarily from the Perry program, do not survive multiplicity adjustment and do not replicate in the other experiments. They also help resolve the discrepancy between the Perry and Abecedarian projects in crime effects. No adult Perry crime effect rejects when controlling FDR at the 5% level, and only one rejects at the 10% level (adult female arrests). It is thus unsurprising that these effects fail to replicate in the Abecedarian study. These facts are noteworthy because much of the Perry program's economic benefits (67%) accrued in the form of reduced crime by participants (Schweinhart, et al. 2005, pp. 148-9). If crime effects are weaker than has been believed, then the often cited 7-to-1 (or greater) benefit-cost ratio for early intervention will be overstated.

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, Schweinhart, Weikart, Marcus, and Xie (2000) report that Head Start participation significantly raises high school graduation rates and lowers arrest rates for females but not males. The results also parallel experimental findings in other areas of the human capital literature. Kling, Liebman, and Katz (2007) 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 services provided under the Job Training Partnership Act (JTPA) significantly increase female earnings at all quantiles, including a 35 percent increase at the lowest quantile. However, JTPA services have no significant effect on males at any quantile below the median, suggesting that disadvantaged males in particular have trouble benefiting from these programs.

In comparison to the ongoing randomized evaluation of Head Start, the three programs discussed in this research demonstrate stronger early effects. Scores on early cognitive tests increase by an average of 0.60 standard deviations in these programs but only 0.14 standard deviations in the Head Start evaluation (U.S. Dept. of Health and Human Services 2005).

However, it is difficult to forecast how these reduced early cognitive effects will affect later life outcomes, and cognitive effects are not reported separately by gender.

6 CONCLUSION

This paper conducts a de novo analysis of the influential early intervention experimental literature using statistical techniques that adjust for multiple inference. It partially confirms previous findings, presenting strong evidence that females benefit from these interventions. Female effects appear in the domains of criminal behavior, marriage, and economic success, but the most consistent improvement is in total years of schooling. These interventions have positive, significant overall long-term effects on females in two of the three programs when adjusting for multiple inference.

For males, however, there is limited evidence of positive long-term treatment effects. Despite several positive and significant (unadjusted) 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 (unadjusted) effects go in both directions and appear at a frequency one would expect simply due to chance. Previous work has missed this finding because there has been no systematic analysis by gender across experiments and because researchers have emphasized the subset of unadjusted significant outcomes rather than applying 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 gives researchers tools to address the issue of multiple testing while minimizing the loss in statistical power. The simulated stand-alone analysis of the most famous (and dramatic) preschool experiment, Perry, demonstrates that applying these tools can generate robust conclusions that are more likely to replicate.

In addition, this paper makes clear several points in the context of the current human capital literature. Foremost, intensive intervention early in life can positively affect later-life outcomes, at least for disadvantaged African-American females. However, there is little evidence of strong long-term 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 these experiments at a time in which advocates are invoking them to support funding for universal preschool education. If treatment effects vary by gender, it is likely that they also vary by race or class. Richer variation in sample demographics is necessary for the design of optimal human capital policy.

APPENDIX A: SUMMARY INDEX DEFINITION

$$\overline{s}_{ij} = \frac{1}{W_{ij}} \sum_{k \in \mathbb{K}_{ij}} w_{jk} \frac{y_{ijk} - \overline{y}_{jk}}{\sigma^{y}_{jk}} \ ,$$

where k indexes outcomes within area j, $\mathbb{K}_{ij}$ is the set of non-missing outcomes for observation i in area j, σ^{y}_{jk} is the control group standard deviation for outcome k in area j, w_{jk} is the outcome weight from the inverted covariance matrix $\hat{\mathbf{\Sigma}}_j^{-\mathbf{1}}$, and $W_{ij} = \sum_{k \in \mathbb{K}_{ij}} w_{jk}$. If K_j is the total number of outcomes for area j, and N_{jmn} is the number of observations not missing for both outcome m and outcome n in area j, then

$$w_{jk} = \sum_{l=1}^{K_j} c_{jkl} \ ,$$

$$\hat{\mathbf{\Sigma}}_j^{-\mathbf{1}} = \begin{bmatrix} c_{j11} & c_{j12} & \cdots & c_{j1K} \\ c_{j21} & c_{j22} & \cdots & \cdots \\ \vdots & \vdots & \ddots & \ddots \\ c_{jK1} & \vdots & \ddots & c_{jKK} \end{bmatrix} \ ,$$

and $\hat{\mathbf{\Sigma}}_j$ consists of elements

$$\hat{\Sigma}_{jmn} = \sum_{i=1}^{N_{jmn}} \frac{y_{ijm} - \overline{y}_{jm}}{\sigma^y_{jm}} \frac{y_{ijn} - \overline{y}_{jn}}{\sigma^y_{jn}} \ .$$

APPENDIX B: POTENTIAL 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 affects all three 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. I therefore impute missing values for key outcomes and examine "worst case" scenarios. Under reasonable assumptions, the paper's central conclusions are unchanged.

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 standard errors that adjust for clustering at the class-by-year level or at the family level. These adjustments do not substantially affect key results.

References

Abadie, A., Angrist, J., and Imbens, G. (2002), "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91–117.

Anderson, M. (2006), "Uncovering Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," Manuscript, Massachusetts Institute of Technology, Dept. of Economics.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the False Discovery Rate," *Journal of the Royal Statistical Society: Series B*, 57, 289–300.

Benjamini, Y., Krieger, A., and Yekutieli, D. (2006), "Adaptive Linear Step-up Procedures that Control the False Discovery Rate," *Biometrika*, 93, 491–507.

Benjamini, Y. and Yekutieli, D. (2001), "The Control of the False Discovery Rate in Multiple Testing Under Dependency," *The Annals of Statistics*, 29, 1165–1188.

Campbell, F. and Ramey, C. (1994), "Effects of Early Intervention on Intellectual and Academic Achievement," *Child Development*, 65, 684–698.

——— (1995), "Cognitive and School Outcomes for High-Risk African-American Students at Middle Adolescence," *American Educational Research Journal*, 32, 743–772.

Campbell, F., Ramey, C., Pungello, E., Sparling, J., and Miller-Johnson, S. (2002), "Early Childhood Education: Young Adult Outcomes From the Abecedarian Project," *Applied Developmental Science*, 6, 42–57.

Carneiro, P. and Heckman, J. (2003), "Human Capital Policy," in *Inequality in America: What Role for Human Capital Policies?*, ed. B. Friedman, Cambridge, MA: MIT Press, pp. 77–240.

Clarke, S. and Campbell, F. (1998), "Can Intervention Early Prevent Crime Later? The

Abecedarian Project Compared with Other Programs," *Early Childhood Research Quarterly*, 13, 319–343.

Currie, J. (2001), "Early Childhood Education Programs," *Journal of Economic Perspectives*, 15, 213–238.

Currie, J. and Thomas, D. (1995), "Does Head Start Make a Difference?," *American Economic Review*, 85, 341–364.

Efron, B. and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, Boca Raton, FL: CRC Press.

Fisher, R. (1935), *The Design of Experiments*, Edinburgh, UK: Oliver Boyd.

Gray, S., Ramsey, B., and Klaus, R. (1982), *From 3 to 20: The Early Training Project*, Baltimore, MD: University Park Press.

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

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

Hill, C., Bloom, H., Black, A., and Lipsey, M. (2007), "Empirical Benchmarks for Interpreting Effect Sizes in Research," MDRC Working Papers on Research Methodology.

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

Kirp, D. (2005), "All My Children," *The New York Times*, July 31, 2005, 20.

Kling, J., Liebman, J., and Katz, L. (2007), "Experimental Analysis of Neighborhood Effects," *Econometrica*, 75, 83–119.

Krueger, A. (2003), "Inequality, Too Much of a Good Thing," in *Inequality in America: What Role for Human Capital Policies?*, ed. B. Friedman, Cambridge, MA: MIT Press, pp. 1–76.

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

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

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

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

Romano, J. and Wolf, M. (2005), "Stepwise Multiple Testing As Formalized Data Snooping," *Econometrica*, 73, 1237–1282.

Rosenbaum, P. (2007), "Interference Between Units in Randomized Experiments," *Journal of the American Statistical Association*, 102, 191–200.

Schweinhart, L., Barnes, H., Weikart, D., Barnett, W. S., and Epstein, A. (1993), *Significant Benefits: The High/Scope Perry Preschool Study Through Age 27*, Ypsilanti, MI: High/Scope Press.

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

Simon, J. (1997), *Resampling: The New Statistics*, Arlington, VA: Resampling Stats.

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

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

U.S. Department of Health and Human Services (2005), "Head Start Impact Study: First Year Findings," Washington, DC.

Westfall, P., Tobias, R., Rom, D., Wolfinger, R., and Hochberg, Y. (1999), *Multiple Comparisons and Multiple Tests Using SAS*, Cary, NC: SAS Institute.

Westfall, P. and Young, S. (1993), *Resampling-Based Multiple Testing*, New York, NY: J. Wiley and Sons.

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test For Heteroskedasticity," *Econometrica*, 48, 817–38.

Williams, V., Jones, L., and Tukey, J. (1999), "Controlling Error in Multiple Comparisons, with Examples from State-to-State Differences in Educational Achievement," *Journal of Educational and Behavioral Statistics*, 24, 42–69.

Figure 1: Effects of Preschool on Teen and Adult Outcomes

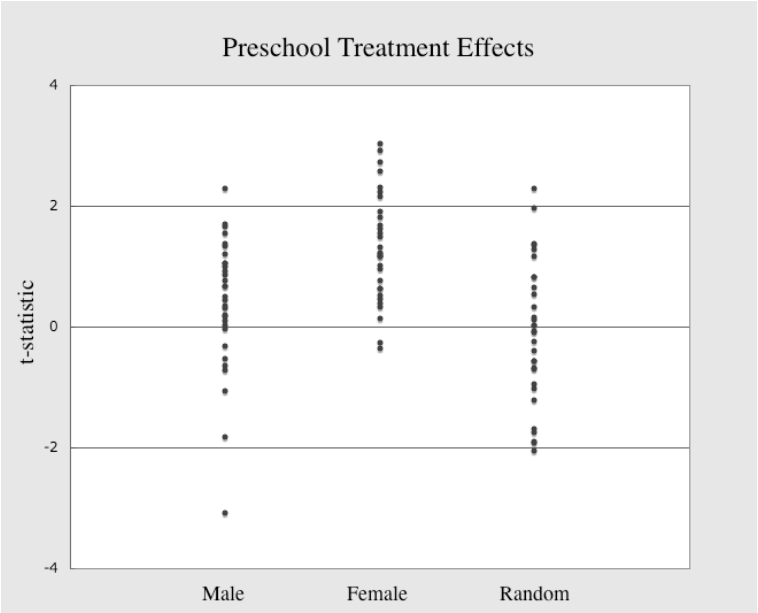

Table 1: Summary Statistics

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)	

NOTE: Parentheses contain standard deviations.

Table 2: Summary Index Components

Project	Stage	Summary Index Components
ABC	Preteen	IQ (5, 6.5, 12), Retained in Grade (12), Special Education (12)
Perry	Preteen	IQ (5, 6, 10), Repeat Grade (17), Special Education (17)
ETP	Preteen	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)

NOTE: Age of measurement in parentheses. For Perry and Early Training grade repetition and special education variables, it was not possible to isolate pre-9th grade outcomes in the data.

Table 3: Summary Index Effects

		Female				**Male**				**Gender**
			Naive	FWER			Naive	FWER		**Diff**
Project	Age	**Effect**	*p*-val	*p*-val	N	**Effect**	*p*-val	*p*-val	N	*t*-stat
ABC	Preteen	0.445	0.026	0.125	54	0.417	0.026	0.184	51	0.11
		(0.194)				(0.181)				
Perry	Preteen	0.537	0.004	0.028	51	0.150	0.387	0.943	72	1.53
		(0.177)				(0.172)				
ETP	Preteen	0.362	0.160	0.349	30	0.148	0.552	0.958	34	0.61
		(0.251)				(0.245)				
ABC	Teen	0.422	0.042	0.156	53	0.162	0.407	0.943	51	0.93
		(0.202)				(0.194)				
Perry	Teen	0.613	0.000	0.003	51	0.035	0.716	0.977	72	3.32
		(0.156)				(0.096)				
ETP	Teen	0.456	0.138	0.349	29	0.123	0.747	0.977	32	0.68
		(0.299)				(0.377)				
ABC	Adult	0.452	0.003	0.024	53	0.312	0.066	0.372	51	0.64
		(0.144)				(0.166)				
Perry	Adult	0.353	0.022	0.125	51	-0.012	0.927	0.977	72	1.83
		(0.150)				(0.130)				
ETP	Adult	-0.069	0.714	0.701	29	-0.710	0.011	0.090	31	1.98
		(0.186)				(0.260)				

NOTE: Parentheses contain OLS standard errors. Naive *p*-values are unadjusted *p*-values based on the *t*-distribution. FWER *p*-values adjust for multiple testing at the summary index level and are computed as described in Section (3.2.2). *t*-statistics test the difference between female and male treatment effects. See Table 2 for the components of each summary index.

Table 4: Effects on Preteen IQ Scores

			Female					**Male**					**Gender Diff**
Outcome	Age	Project	**Effect**	CM	Naive *p*-val	FDR *q*-val	N	**Effect**	CM	Naive *p*-val	FDR *q*-val	N	*t*-stat
IQ	5	ABC	4.94	96.76	0.176	0.304	48	10.19	90.81	0.005	0.082	47	-1.05
			(3.58)					(3.52)					
IQ	6.5	ABC	5.13	92.96	0.134	0.271	46	7.18	92.10	0.053	0.517	45	-0.41
			(3.35)					(3.65)					
IQ	12	ABC	8.35	87.35	0.004	0.048	52	3.21	90.48	0.294	1.000	49	1.24
			(2.75)					(3.10)					
IQ	5	Perry	12.67	81.65	0.004	0.048	39	10.61	84.79	0.001	0.049	54	0.40
			(4.30)					(2.84)					
IQ	6	Perry	3.75	87.16	0.241	0.318	48	5.66	85.82	0.037	0.451	72	-0.46
			(3.21)					(2.68)					
IQ	10	Perry	4.96	81.79	0.173	0.304	43	-2.33	86.03	0.372	1.000	71	1.70
			(3.45)					(2.56)					
IQ	5	ETP	13.55	87.60	0.015	0.077	30	4.43	87.18	0.232	1.000	34	1.28
			(6.09)					(3.75)					
IQ	7	ETP	8.61	89.89	0.118	0.271	29	4.11	92.89	0.344	1.000	30	0.57
			(6.69)					(4.25)					
IQ	10	ETP	9.79	81.56	0.067	0.216	29	-3.17	88.33	0.511	1.000	27	1.68
			(5.73)					(5.15)					

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

Table 5: Effects on Preteen Primary School Outcomes

			Female					**Male**					**Gender Diff**
Outcome	Age	Project	**Effect**	CM	Naive *p*-val	FDR *q*-val	N	**Effect**	CM	Naive *p*-val	FDR *q*-val	N	*t*-stat
Retained	12	ABC	-0.229	0.429	0.080	0.216	53	-0.188	0.545	0.197	1.000	50	-0.21
			(0.125)					(0.142)					
Spec Educ	12	ABC	-0.066	0.296	0.567	0.453	53	-0.269	0.591	0.057	0.517	50	1.10
			(0.123)					(0.140)					
Repeat Grade	12	Perry	-0.201	0.409	0.133	0.271	46	0.078	0.389	0.520	1.000	66	-1.51
			(0.137)					(0.124)					
Spec Educ	17	Perry	-0.262	0.462	0.061	0.216	51	-0.037	0.462	0.733	1.000	72	-1.28
			(0.129)					(0.119)					
Retained	17	ETP	-0.284	0.600	0.154	0.290	29	0.100	0.600	0.552	1.000	30	-1.40
			(0.195)					(0.192)					
Special Help	17	ETP	0.116	0.200	0.504	0.446	29	0.036	0.364	0.817	1.000	31	0.31
			(0.171)					(0.188)					

NOTE: Parentheses contain robust standard errors. CM refers to control mean. Sample size varies within experiments due to attrition for some variables. *p*- and *q*-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**
					Naive	FDR				Naive	FDR		**Diff**
Outcome	Age	Project	**Effect**	CM	*p*-val	*q*-val	N	**Effect**	CM	*p*-val	*q*-val	N	*t*-stat
IQ	15	ABC	4.22	89.50	0.144	0.281	53	4.66	92.48	0.094	0.674	51	-0.11
			(2.85)					(2.79)					
IQ	14	Perry	2.64	76.77	0.311	0.359	46	-0.96	83.26	0.755	1.000	64	0.91
			(2.57)					(3.03)					
IQ	17	ETP	2.08	76.11	0.739	0.524	25	1.64	76.78	0.741	1.000	28	0.05
			(6.80)					(5.09)					
HS Grad	18	ABC	0.226	0.607	0.081	0.216	52	-0.096	0.739	0.468	1.000	51	1.80
			(0.122)					(0.131)					
HS Grad	18	Perry	0.494	0.346	0.000	0.001	51	-0.061	0.667	0.575	1.000	72	3.32
			(0.121)					(0.115)					
Ever Drop	18	ETP	-0.289	0.500	0.101	0.245	29	-0.095	0.545	0.654	1.000	31	-0.72
Out of HS			(0.190)					(0.193)					

NOTE: Parentheses contain robust standard errors. CM refers to control mean. Sample size varies within experiments due to attrition for some variables. *p*- and *q*-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**
					Naive	FDR				Naive	FDR		**Diff**
Outcome	Age	Project	**Effect**	CM	*p*-val	*q*-val	N	**Effect**	CM	*p*-val	*q*-val	N	*t*-stat
Unemp	19	Perry	-0.308	0.708	0.027	0.111	49	-0.021	0.385	0.877	1.000	72	-1.60
			(0.138)					(0.116)					
Transfers	19	Perry	-1,569	2,828	0.035	0.134	51	-28	398	0.936	1.000	72	-1.96
			(722)					(319)					
Ever Work	18	ETP	0.125	0.500	0.591	0.453	22	-0.063	1.000	0.674	1.000	23	0.73
			(0.249)					(0.063)					
Teen Parent	19	ABC	-0.211	0.571	0.125	0.271	53	-0.126	0.304	0.325	1.000	51	-0.47
			(0.137)					(0.123)					
Had Child	19	Perry	-0.187	0.667	0.205	0.304	49	-0.044	0.256	0.665	1.000	72	-0.82
			(0.142)					(0.101)					
Arrested	19	Perry	-0.337	0.417	0.005	0.048	49	-0.079	0.564	0.550	1.000	72	-1.54
			(0.117)					(0.119)					

NOTE: Parentheses contain robust standard errors. CM refers to control mean. Sample size varies within experiments due to attrition for some variables. *p*- and *q*-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 Diff**
Outcome	Age	Project	**Effect**	CM	Naive *p*-val	FDR *q*-val	N	**Effect**	CM	Naive *p*-val	FDR *q*-val	N	*t*-stat
In College	21	ABC	0.293	0.107	0.016	0.077	53	0.148	0.174	0.267	1.000	51	0.87
			(0.116)					(0.121)					
Any College	27	Perry	0.160	0.280	0.260	0.336	50	-0.005	0.308	0.971	1.000	72	0.94
			(0.137)					(0.110)					
In Post HS Educ	21	ETP	0.121	0.300	0.524	0.453	29	-0.486	0.636	0.004	0.082	31	2.37
			(0.191)					(0.171)					

NOTE: Parentheses contain robust standard errors. CM refers to control mean. Sample size varies within experiments due to attrition for some variables. *p*- and *q*-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
					Naive	FDR				Naive	FDR		**Diff**
Outcome	Age	Project	**Effect**	CM	*p*-val	*q*-val	N	**Effect**	CM	*p*-val	*q*-val	N	*t*-stat
Employed	21	ABC	0.104	0.536	0.427	0.405	53	0.188	0.455	0.199	1.000	50	-0.43
			(0.137)					(0.142)					
Employed	27	Perry	0.255	0.545	0.078	0.216	47	0.036	0.564	0.773	1.000	69	1.20
			(0.136)					(0.121)					
Annual Income	27	Perry	2,567	8,986	0.347	0.390	47	2,363	12,495	0.391	1.000	66	0.05
			(2,686)					(2,708)					
Monthly Income	27	Perry	396	651	0.101	0.245	47	537	830	0.026	0.388	68	-0.41
			(236)					(247)					
Employed	40	Perry	0.015	0.818	0.931	0.574	46	0.200	0.500	0.112	0.741	66	-1.12
			(0.115)					(0.120)					
Annual Income	40	Perry	3,492	17,374	0.538	0.453	46	6,228	21,119	0.299	1.000	66	-0.34
			(5,491)					(5,958)					
Monthly Income	40	Perry	162	1,615	0.704	0.505	46	436	1,839	0.459	1.000	66	-0.39
			(431)					(562)					
Receive Income	21	ETP	-0.074	0.600	0.697	0.505	29	-0.159	0.909	0.304	1.000	31	0.36
			(0.200)					(0.134)					
Receive Welfare	21	ETP	-0.042	0.200	0.826	0.537	30	N/A					
			(0.157)										

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

Table 10: Effects on Adult Social Outcomes

			Female					**Male**					**Gender Diff**
Outcome	Age	Project	**Effect**	CM	Naive *p*-val	FDR *q*-val	N	**Effect**	CM	Naive *p*-val	FDR *q*-val	N	*t*-stat
Convicted	21	ABC	-0.101	0.143	0.240	0.318	52	-0.089	0.348	0.532	1.000	50	-0.08
			(0.079)					(0.133)					
Felony	21	ABC	N/A					-0.113	0.261	0.364	1.000	50	
								(0.117)					
Jailed	21	ABC	-0.030	0.071	0.761	0.529	52	-0.177	0.391	0.165	1.000	51	1.01
			(0.065)					(0.131)					
Marijuana User	21	ABC	-0.317	0.357	0.003	0.048	53	-0.127	0.435	0.376	1.000	49	-1.10
			(0.101)					(0.140)					
Criminal Record	27	Perry	-0.146	0.346	0.268	0.336	51	-0.021	0.718	0.828	1.000	72	-0.75
			(0.125)					(0.109)					
Lifetime Arrests	27	Perry	-1.95	2.27	0.011	0.069	49	-2.31	6.10	0.126	0.771	72	0.21
			(0.83)					(1.50)					
Ever Used Drugs	27	Perry	-0.157	0.300	0.213	0.304	41	0.198	0.189	0.070	0.560	68	-2.08
			(0.131)					(0.110)					
Married	27	Perry	0.317	0.083	0.009	0.066	49	0.002	0.256	0.969	1.000	70	2.01
			(0.115)					(0.107)					

NOTE: Parentheses contain robust standard errors. CM refers to control mean. Sample size varies within experiments due to attrition for some variables. *p*- and *q*-values are computed as described in Section 3; *t*-statistics test the difference between female and male treatment effects. No female in the Abecedarian treatment or control group was arrested for a felony.

Figure Caption:

Figure 1: *t*-statistics for teen and adult outcomes. Each point is a *t*-statistic for a single outcome, and the positive direction corresponds to a "better" outcome. The first column plots male *t*-statistics, the second column plots female *t*-statistics, and the third column plots a set of randomly generated *t*-statistics.