Abstract Title Page
Not included in page count.

Title: Do the Effects of Early Childhood Programs on Academic Outcomes Vary by Gender? A Meta-Analysis

Author(s): Robert Kelchen, University of Wisconsin-Madison (kelchen@wisc.edu),
Katherine Magnuson, University of Wisconsin-Madison (kmagnuson@wisc.edu),
Greg Duncan, University of California-Irvine (gduncan@uci.edu),
Holly Schindler, Center for the Developing Child, Harvard University (holly_schindler@harvard.edu),
Hilary Shager, University of Wisconsin-Madison (hshager@wisc.edu), &
Hiro Yoshikawa, Harvard Graduate School of Education (yoshikhi@gse.harvard.edu).

Abstract Body

Limit 5 pages single spaced.

Background and Motivation:

It has become typical for children to attend some type of early childhood education (ECE) before entering kindergarten. This reflects both a greater attention to learning in the early years, as well as mothers' increased participation in the formal labor market (Magnuson, Meyers, & Waldfogel, 2007). For decades, scholars, policy-makers, and advocates have touted the potential of ECE to remediate disadvantaged children's low levels of achievement at school entry, and have more recently made the argument that these programs may also be beneficial for more affluent children. Although there has been a proliferation of evaluations of early education programs, the argument that these programs have lasting effects, particularly for economically disadvantaged children, has been largely based on a few early, small, high quality experimental studies. Most prominent among these studies has been the evaluation of the Perry Preschool Program, which found that a year or two of high quality early education boosted children's early IQ and achievement skills, as well as their later school attainment and earnings.

A recent reanalysis of Perry Preschool and two other prominent experimental ECE studies (Abecedarian and the Early Training Project) by Anderson (2008) comes to a provocative conclusion, finding that female participants gained substantially from the programs, but "the overall patterns of male coefficients is consistent with the hypothesis of minimal effects at best--significant (unadjusted) effects go in both directions and appear at a frequency that would be expected due simply to chance" (Anderson, 2008, p. 1494). He found that although males in Abecedarian and Perry Preschool had early gains in IQ measures, these effects were neither found in later years nor in other outcomes, such as special education placement or grade retention. The Early Training Project demonstrated no significant benefits for boys' IQ even at program completion, although positive effects were found for girls.

Anderson's work raises the question of whether the presumed benefits of public and private investments in ECE are as broad as previously assumed. Is the finding that boys do not demonstrate as large or as long-lasting educational gains from early childhood programs endemic to all early childhood education programs? Or is something particular to the set of studies Anderson analyzed? Although boys are thought to be more sensitive to environmental contexts and to be less developmentally advanced than girls, at least in early childhood (Crockenberg, 2003; Zaslow & Haynes, 1986), the implications of the theoretical literature on the effects of early education by gender is unclear. If early education programs are designed to be compensatory; i.e., boosting the skills of low-performing students, then boys should benefit more than girls. If, however, early education programs foster a "skills beget skills" learning process, then girls, who are more likely to have stronger basic skills, should benefit more than boys.

Identifying whether differential gender impacts exist more broadly in early education seems particularly important given recent achievement data suggesting that girls consistently outperform boys on both the NAEP reading and math tests, and also have higher levels of attainment than boys (Aud, et al., 2010). Is the capacity of girls to benefit more from ECE programming one reason why?

Using data on a larger and more representative set of ECE evaluations, and rigorous meta-analytic methods, this paper will investigate whether ECE programs have differential effects on boys and girls in three domains representing cognitive skills, academic achievement, and other school-related outcomes.

Research Question:

Does the impact of ECE programs on the cognitive, achievement, and other school-related outcomes of students differ by gender? We will further consider whether any such effects differ by the domain of outcome (cognitive, achievement, or other school outcomes such as grade retention and special education placement), and the timing of the outcome measurement (at program completion or a later follow-up).

Research Design:

Meta-analysis. To understand whether the effects of ECE programs differ by gender, we will conduct a meta-analysis, a method of quantitative research synthesis that uses prior study results as the unit of observation (Cooper & Hedges, 2009). To combine findings across studies, estimates are transformed into a common metric called an "effect size," expressed as a fraction of a standard deviation. Outcomes from individual studies can then be used to estimate the average effect size across studies. Additionally, meta-analysis can be used to test whether average effect size differs by characteristics of the studies (e.g. gender of participants). After defining the problem of interest, meta-analysis proceeds in the following steps, described below: 1) literature search, 2) data evaluation, and 3) data analysis.

Literature Search. The ECE studies analyzed in this paper compose a sub-set of studies from a large meta-analytic database being compiled by The National Forum on Early Childhood Program Evaluation. This database includes studies of child and family policies, interventions, and prevention programs provided to children from the prenatal period to age five, building on a previous meta-analytic database created by Abt Associates, Inc. (Jacob, Creps, & Boulay, 2004; Layzer, Goodson, Bernstein, & Price, 2001).

The original Abt database contained 107 ECE programs evaluated between 1960 and 2003, but this database did not include all potential ECE evaluations for three- to five-year olds. We used a number of search strategies to identify as many published and unpublished program evaluations conducted between 1960 and 2007 that met our programmatic and methodological criteria for inclusion. First, we conducted keyword searches in the ERIC, PsychINFO, and Dissertation Abstracts databases. Next, the research team tracked down additional reports mentioned in collected studies. Our research team then searched additional specialized databases, government databases, ECE policy group websites, and conference programs as well as contacting researchers in the field. Over 200 new ECE evaluations were identified, in addition to the approximately 73 originally coded by Abt that met our general screening criteria.

Data Evaluation. The next step in the meta-analysis process is to determine whether identified studies meet our established inclusion criteria: studies must have i) a comparison group (either an observed control or alternative treatment group); and ii) at least ten participants in each condition, with attrition of less than 50 percent.[2] Evaluations may be experimental or quasi-experimental, using one of the following designs: regression discontinuity, fixed effects (individual or family), difference in difference, instrumental variables, propensity score matching, or interrupted time series. Quasi-experimental evaluations not using one of the former analytic strategies are also screened in if they include a comparison group *plus* pre-and post-test

[2] Because some of our inclusion criteria differed from Abt's original criteria, we re-screened all of the studies included in the original database as well as the new ones identified by the Forum research team.

information on the outcome of interest or demonstrate adequate comparability of groups on baseline characteristics (determined by a joint test).

For this particular study, which is focused on comparing the effects of center-based ECE programs by gender, we impose some additional inclusion criteria. We include only studies that measure differences between center-based ECE participants and control groups that were assigned to receive no equivalent services.[3] For example, studies that compared the effects of Head Start to another type of early education program or examined a curricular add-on in pre-kindergarten are excluded. We then exclude all studies (and outcomes within studies) that did not provide analyses of programs' results separately by gender. In addition, we include only studies that provide at least one measure of children's cognitive, achievement, or other school-related outcomes. (See Appendix A2 for a list of all articles that met our inclusion criteria.)

Coding Studies. For reports that met our inclusion criteria, the research team developed a protocol to codify information about study design, program and sample characteristics, as well as statistical information needed to compute effect sizes. This protocol serves as the template for the database and delineates all the information about an evaluation that we want to describe and analyze. A team of a dozen graduate research assistants were trained as coders during a 3- to 6-month process that included instruction in evaluation methods, using the coding protocol, and computing effect sizes. Before coding independently, research assistants also passed a reliability test. Questions about coding were resolved in weekly research team conference calls.

Database. The resulting database is organized in a three-level hierarchy (from highest to lowest): the program, the contrast, and the effect size. A "program" is defined as a collection of comparisons in which the treatment group received a particular model of center-based ECE and is compared to another sample of children drawn from the same sample pool who received no equivalent services. One ECE report included evaluations of four programs, and these are considered separate programs in our data. Each program also produces a number of "contrasts," defined as a comparison between one subsample of children who received center-based ECE and another subsample of children who received no equivalent services. Programs included in our study have at least two contrasts—one for boys and one for girls--nested within one program.

The data for this study include 20 ECE programs and 68 contrasts, 34 each for boys and girls. In turn, within each contrast there are multiple individual "effect sizes" (estimated standard deviation unit difference in an outcome between the children who experienced center-based ECE and those who did not), corresponding to the particular measures that are used. The 68 contrasts in the database provide a total of 582 effect sizes.[4] The average posttest sample size for the treatment and control groups is 65 and 53 children, respectively. (See Table 1: Key Meta-Analysis Terms and Sample Sizes.)

Effect size computation. Outcome information was reported using a number of different statistics, which were converted to effect sizes (Hedges' g) with the commercially available software package Comprehensive Meta-Analysis (Borenstein, Hedges, Higgins, & Rothstein,

[3] However, studies are not excluded if children assigned to a no alternative treatment control group sought services of their own volition.

[4] In several studies, outcomes were mentioned in the text, but not enough information was provided to calculate effect sizes; for example, references were made to non-significant findings, but no numbers were reported. There are 142 effect sizes within ten programs with at least some missing information; the non-missing sample consists of 440 effect sizes within 17 programs. Excluding such effect sizes could lead to upward bias of treatment effects; therefore, we coded all available information for such measures, but coded actual effect sizes as missing. We will test the sensitivity of our findings by assigning missing effect sizes a range of plausible values.

2005). Hedges' g is an effect size statistic that makes an adjustment to the standardized mean difference (Cohen's d) to account for bias in the d estimator when sample sizes are small. 58 of the 68 contrasts provided more than one effect size to the analysis.

Measures. The dependent variables in these analyses are the effect sizes measuring the impact of ECE on children's cognitive skills, achievement, and other school-related outcomes. The cognitive outcomes include measures of IQ, vocabulary, theory of mind, attention, task persistence, and syllabic segmentation, such as rhyming. Achievement outcomes include measures of reading, math, letter recognition, and numeracy skills. School-related outcomes encompass attendance, grades received, retention, special education, educational aspirations, and attainment. Currently coded effect sizes range from -1.04 to 1.59, with an average weighted effect size of .19.

Due to the balanced nature of the dataset (boys and girls experienced the same programs and were given the same tests) and the way in which we conduct our analysis (described in more detail below), there is little need to control for differences in program characteristics in the statistical analysis. However, for descriptive purposes, characteristics of the ECE programs are presented in Table 2. (See Table 2: Summary Statistics of the Meta-Analytic Dataset.)

Statistical analysis. Our key research question is whether the effect of ECE programs on the cognitive, achievement, and school-related outcomes of children differs by gender. To test this hypothesis requires a multivariate, multi-level approach to modeling these associations. The level-1 effect size model is:

$$(1)\ ES_{ij} = \beta_{0i} + \beta_{1i}x_{1ij} + \beta_{2i}x_{2ij} + e_{ij}$$

In this equation, each effect size (ES_{ij}), for program *i* and effect size *j*, is modeled as a function of the intercept (β_{0i}), which represents the average effect size among all programs, the key parameter of interest--a dummy variable for whether the effect size is for all boys or all girls ($\beta_{1i}x_{1ij}$), a small number of covariates measuring features of the effect sizes such as domain of the outcome and timing of the outcome ($\beta_{2i}x_{2ij}$), and a within-program error term (e_{ij}). The level-2 equation (program level) models the intercept as a function of the grand mean effect size for the program (β_0) and a between-program random error term (u_i):

$$(2)\ \beta_{0i} = \beta_0 + u_i$$

This "mixed effects" model assumes that there are two sources of variation in the effect size distribution, beyond subject-level sampling error: 1) the "fixed" effects of between-effect size variables measured by gender and other effect size covariates; and 2) remaining "random" unmeasured sources of variation between and within programs. To account for differences in precision of effect size estimates as well as the difference in the number of estimates provided by each program, regressions are weighted by the inverse variance weight of each effect size multiplied by the inverse of the number of effect sizes within a program (Lipsey & Wilson, 2001).

Supplementary models will be estimated to consider how this main effect of gender may differ by the domain of the outcome and the timing of the outcome assessment. We will estimate separate models for each outcome domain, and also estimate interaction terms (gender by outcome domain) to test differences. Likewise, we will estimate separate models as well as

include interaction terms to test for effect sizes measured at or shortly after program completion and those measured at later points.

Additional Data. Thus far, the screening process, based on the above criteria, has resulted in the inclusion of 17 ECE publications or reports representing 20 different interventions. There are still approximately 30-40 publications to be coded; if the current rate of finding programs which meet the inclusion criteria continues, we expect a few additional programs to be added to the database. All ECE coding in the database is expected to be complete by the end of 2010.

Data Limitations. At this point in time, there are two major limitations with the data. The majority of the current effect sizes are from programs that began before 1972, although any additional effect sizes that will be added to the dataset will likely be from more recent studies. While this may limit our ability to generalize our findings to more recent cohorts of children and programs, it is important to recognize that Anderson's analysis also relied on older studies, and these studies are the only source of long-term outcome data. Second, most of the effect sizes are from the cognitive domain, rather than the achievement or school-related outcomes domains. Although we will consider each domain separately, there will be less power to detect effects separately by domain.

Findings / Results:

Description of the main findings with specific details.

In a preliminary analysis, we tested for differential program effects by gender with the outcomes of the three domains combined (with no other effect size or program level covariates). Our results suggest that the difference between boys and girls on this broad academic outcome is small and non-significant (effect size of .010 favoring girls, p-value=.611). This suggests that center-based ECE programs have similar overall effects on boys and girls; however, it may be the case that gender differences may exist by domain and the timing of the outcome. Thus, we will we also investigate whether gender differences emerge in particular outcome domains or by the timing of the outcome measures.

Conclusions:

Our preliminary findings of achievement, cognitive, and other school-related outcomes measured in evaluations of early childhood programs suggest that the broad achievement and school outcomes of boys and girls are quite similar, and that both genders benefit by approximately two-tenths of a standard deviation on the average outcome. Indeed, the difference is not only statistically insignificant, but also substantively minimal. This conclusion differs from Anderson's broad conclusions about boys experiencing minimal gains from ECE programs. Indeed, our data suggest that boys gain as much from ECE as girls, at least on the range of outcomes available in our data. It will be important to determine whether this pattern of findings holds up across each outcome domain and regardless of the timing of the outcome measures.

If these findings are robust to our future alternative specifications, it suggests that early education programs neither exacerbate nor remediate any early gender advantages in cognitive and other achievement outcomes. Thus, if reducing gender disparities is a worthy educational goal, other policies and practices will need to be considered.

Appendices

Not included in page count.

Appendix A1. References in the Abstract

References are to be in APA version 6 format.

Anderson, M. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training projects. *Journal of the American Statistical Association , 103* (484), 1481-1495.

Aud, S., Hussar, W., Planty, M., Snyder, T., Bianco, K., Fox, M. A., et al. (2010). *The condition of education 2010*. Institute of Education Sciences, National Center for Education Statistics. Washington, DC: U.S. Department of Education.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). *Comprehensive meta-analysis, version 2.* Englewood, New Jersey: Biostat.

Christian, K., Morrison, F., Frazier, J., & Massetti, G. (2000). Specificity in the nature and timing of cognitive growth in kindergarten and first grade. *Journal of Cognition and Development , 1* (4), 429-448.

Cooper, H., & Hedges, L. V. (2009). Research synthesis as a scientific process. In H. Cooper, L. V. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 3-17). New York, New York: Russell Sage Foundation.

Crockenberg, S. C. (2003). Rescuing the baby from the bathwater: How gender and temperament may influence how child care affects child development. *Child Development, 74*, 1034-1038.

Jacob, R., Creps, C., & Boulay, B. (2004). *Meta-analysis of research and evaluation studies in early childhood education.* Cambridge, Massachusetts: Abt Associates, Inc.

Layzer, J., Goodson, B., Bernstein, L., & Price, C. (2001). *National evaluation of family support programs, volume A: The meta-analysis, final report.* Cambridge, Massachusetts: Abt Associates, Inc.

Lipsey, M., & Wilson, D. (2001). *Practical meta-analysis.* Thousand Oaks, California: Sage Publications.

Magnuson, K. A., Meyers, M. K., & Waldfogel, J. (2007). Public funding and enrollment in formal child care in the 1990s. *Social Science Review , 81* (1), 47-83.

Royston, P. (2004). Multiple imputation of missing values. *Stata Journal , 4*, 227-241.

Zaslow, M. S., & Hayes, C. D. (1986). Sex differences in children's responses to psychosocial stress: Toward a cross-context analysis. In M. Lamb & B. Rogoff (Eds.), *Advances in developmental psychology* (Vol. 4, pp. 289–337). Hillsdale, NJ: Erlbaum.

Appendix A2. References for Articles Containing Effect Sizes in the Analysis

Anderson, M. L. (2006). *Essays in public health and early education.* Cambridge, Massachusetts: ProQuest Information and Learning.

Appalachian Educational Laboratory. (1970). *Evaluation report: Early childhood education program, 1969 field test.* Charleston, West Virginia: Appalachian Educational Laboratory.

Belfield, C. R., Nores, M., Barnett, S., & Schweinhart, L. (2006). The High/Scope Perry Preschool program: Cost-benefit analysis using data from the age-40 followup. *The Journal of Human Resources , 41* (1), 162-190.

Campbell, F. A., & Pungello, E. (2000). High-quality child care has long-term educational benefits for poor children. *Paper presented at the Fifth Head Start National Research Conference.* Washington, DC.

Capobianco, R. J. (1967). A pilot project for culturally disadvantaged preschool children. *Journal of Special Education , 1*, 191-194.

Clarke, S. H., & Campbell, F. A. (1998). Can intervention early prevent crime later? The Abecedarian Project compared with other programs. *Early Childhood Research Quarterly , 13* (2), 319-343.

Coffman, A. O., & Dunlap, J. M. (1968). *The effects of assessment and personalized programming on subsequent intellectual development of prekindergarten and kindergarten children.* University City, Missouri: School District of University City.

Di Lorenzo, L. T., Salter, R., & Brady, J. J. (1969). *Prekindergarten programs for educationally disadvantaged children.* New York, New York: New York State Department of Education.

Esteban, M. (1987). *A comparison of Head Start and non-Head Start reading readiness scores of low-income kindergarten children of Guam.* Ann Arbor, Michigan: UMI Dissertation Services.

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

Herzog, E., Newcomb, C., & Cisin, I. (1972). *Preschool and postscript: An evaluation of an inner-city program.* Washington, DC: Social Research Group.

Hines, B. W. (1971a). *Analysis of Intellgigence Scores.* Charleston, West Virginia: Appalachian Educational Laboratory.

Hines, B. W. (1971b). *Analysis of visual perception of children in the Appalachia Preschool Education Program.* Charleston, West Virginia: Appalachia Educational Laboratory.

Hines, B. W. (1971c). *Attainment of cognitive objectives.* Charleston, West Virginia: Appalachia Educational Laboratory.

Hines, B. W. (1971d). *Detailed analysis of the language development of children in AEL's Preschool Education Program.* Charleston, West Virginia: Appalachia Educational Laboratory.

Jackson, R., McCoy, A., Pistorino, C., Wilkinson, A., Burghardt, J., Clark, M., et al. (2007). *National evaluation of Early Reading First: Final report.* Institute of Education Sciences, U.S. Department of Education. Washington: Government Printing Office.

Kraft, I., Fuschillo, J., & Herzog, E. (1968). *Prelude to school: An evaluation of an inner-city preschool program.* Children's Bureau, Social and Rehabilitation Service. Washington, DC: Government Printing Office.

Krider, M. A., & Petsche, M. (1967). *An evaluation of Head Start pre-school enrichment programs as they affect the intellectual ability, the social adjustment, and the achievement level of five-year-old children enrolled in Lincoln, Nebraska.* Lincoln, Nebraska: University of Nebraska.

MacDonald, R. (1971). *Analysis of intelligence scores.* Charleston, West Virginia: Appalachia Educational Laboratory.

Miller, L., & Bizzell, R. (1984). Long-term effects of four preschool programs: Ninth- and tenth-grade results. *Child Development , 55* (4), 1570-1587.

Miller, L., Dyer, J., Stevenson, H., & White, S. (1975). Four preschool programs: Their dimensions and effects. *Monographs of the Society for Research in Child Development , 40* (5/6), 1-170.

Montgomery County Public Schools. (1970). *Impact of the Head Start program. Phase I of a projected longitudinal study to the U. S. Office of Economic Opportunity. Final report.* Washington, DC: Office of Economic Opportunity.

Newton, J. (2006). *Special education and school success in third grade: Does Head Start make a difference?* Minneapolis, Minnesota: Walden University.

Porter, P. J., Leodas, C., Godley, R. A., & Budroff, M. (1965). *Evaluation of Headstart educational program in Cambridge, MA.* Cambridge, Massachusetts: Harvard University.

Schweinhart, L., Montie, J., Xiang, Z., Barnett, W., Belfield, C., & Nores, M. (2005). *Lifetime effects: The High/Scope Perry Preschool study through age 40.* Ypsilanti, Michigan: High/Scope Press.

Smith, E. J., Pellin, B. J., & Agruso, S. A. (2003). *Bright Beginnings: An effective literacy-focused preK program for educationally disadvantaged four-year-old children.* Arlington, Virginia: Educational Research Service.

Steglich, W. G., & Cartwright, W. J. (1965). *Report of the effectiveness of Project Head Start, Lubbock, Texas. Parts I, II, and appendices.* Lubbock, Texas: Texas Technological College.

Appendix B. Tables and Figures

Not included in page count.

Table 1: Key Meta-Analysis Terms and Sample Sizes

Term	**Description**	**N in current database***
Report	Written evaluation of early childhood education by gender (e.g., a journal article, government report, book chapter) containing effect sizes and meeting inclusion criteria	27
Program	Collection of comparisons in which groups are assigned to distinct treatment models and control groups	20
Contrast	Comparison between one group of children who received center-based ECE and another group of children who received no equivalent services	68
Effect Size	Measure of the difference in cognitive outcomes between the children who experienced center-based ECE and those who received no equivalent services, expressed in standard deviation units (*Hedges' g*)	582

*Note: We estimate that our database currently contains approximately 85 percent of the studies that will be in the final analysis.

Table 2: Summary Statistics of the Dataset (N=582 effect sizes)[5]

Starting year of program	Missing: 4 1960-1972: 530 1973-2007: 48
Number of sites	Missing: 82 One: 276 Two or more: 224
Urbanicity	Missing or mix: 62 Urban or suburban: 252 Rural: 268
Method of assignment	Random: 114 Quasi-experimental: 404 Post-hoc design change: 64
Length of treatment	Missing: 76 <12 months: 242 13-24 months: 160 25+ months: 104
Other services received by control group?	None: 412 Some: 170
Outcome domain	Cognitive skill: 346 Achievement: 152 Other school outcomes: 84
Months elapsed since end of treatment	Missing: 76 During treatment: 120 0-12 months: 200 13-24 months: 67 25+ months: 119

[5] 142 of the 582 effect sizes have at least some missing data (i.e. the direction of the effect may be known, but not the significance).