Abstract Title Page
Not included in page count.

Title: Using Meta-analysis to Explain Variation in Head Start Research Results: The Role of Research Design

Author(s):
Hilary M. Shager
University of Wisconsin-Madison

Holly S. Schindler
Center on the Developing Child
Harvard University

Cassandra M.D. Hart
Northwestern University

Greg J. Duncan
Department of Education
University of California-Irvine

Katherine A. Magnuson
University of Wisconsin-Madison

Hirokazu Yoshikawa
Harvard Graduate School of Education

Abstract Body

Background and Significance:

Head Start was designed as a holistic intervention to improve economically disadvantaged, preschool-aged children's cognitive and social development by providing a comprehensive set of educational, health, nutritional, and social services, as well as opportunities for parent involvement (Zigler & Valentine, 1979). Since its inception in 1965, the federally funded program has enrolled over 25 million children; yet, despite its longevity, questions regarding Head Start's effectiveness remain (Currie & Thomas, 1995; Nathan, 2007; US GAO, 1997). Evaluations of Head Start vary greatly in method and quality, and although previous reviews have described such differences and compared evaluations in a subjective, narrative manner, there has been little empirical investigation of the importance of such factors in explaining differing results. By taking prior outcome estimates as the unit of study, a meta-analysis provides a unique opportunity to estimate associations between research design characteristics and evaluation results. Our proposed study uses newly-coded information from Head Start impact studies conducted between 1965 and 2007 to explore how research designs are related to variation in measures of the program's impact on children's cognitive and achievement outcomes.

The only existing meta-analysis of Head Start research, conducted over 25 years ago by McKey and colleagues (1985), included studies completed between 1965 and 1982, and found positive program impacts on cognitive test scores in the short term (effect sizes=.31 to .59), but not the long term (two or more years after program completion; effect sizes= -.03 to .13). Initial descriptive analyses of methodological factors such as quality of study design, sampling, and attrition revealed only slight influences on the magnitude and direction of effect sizes; therefore, these variables were not included in the main analyses. The authors found, however, that studies with pre-/post-test designs (which may not adequately control for maturation effects) tended to produce larger effect sizes than treatment/control group studies.

More general meta-analyses of early childhood education (ECE) programs, which include some Head Start studies, have found significant links between study characteristics and results. For example, Camilli, Vargas, Ryan, and Barnett (2008) found that studies with high quality design (measured by a composite of indicators such as attrition and baseline equivalence of treatment and control groups) yielded larger effect sizes for cognitive outcomes. In contrast, two meta-analyses of longitudinal ECE program evaluations did not find a significant link between effect sizes for cognitive outcomes and research design features such as design type, sample size, attrition rate, baseline equivalence, and quality of outcome measures (Gorey, 2001; Nelson, Westhues & MacLeod, 2003).

Research Question:

Given the current interest in ECE as an intervention strategy for disadvantaged children and the magnitude of public investment in Head Start ($6.9 billion in FY 2007), it is important for researchers and policy makers to be effective designers and consumers of Head Start evaluations (Office of Head Start, 2008). Although some previous meta-analyses suggest a link between evaluation characteristics and results, evidence is mixed, and recent methodological advances have not been considered. A more detailed empirical test of the contribution of particular research design characteristics is needed to enable scholars to better understand findings from prior studies, as well as to inform future studies. Our study will investigate the

role of such factors in explaining variation in Head Start evaluation results for children's cognitive and achievement outcomes. Specifically, we test whether the following research design characteristics explain heterogeneity in the estimated effects of Head Start on children's cognitive and achievement outcomes: type and rigor of design, quality of dependent measure, attrition, and activity level of control group. We will also pay attention to the timing of the outcome, distinguishing between effects at program completion and subsequent follow-up assessments.

Because Head Start primarily serves disadvantaged children, the concern with many prior studies of the program is that analysts did not do enough to control for the independent influence such disadvantage might have on outcomes, thus, downwardly biasing estimates of Head Start effectiveness (Currie & Thomas, 1995). Therefore, we hypothesize that studies that use rigorous methods to ensure similarity between treatment and control groups, in terms of baseline characteristics, will produce larger effect sizes. We also expect that higher quality outcome measures, which introduce less measurement error, will be associated with larger average effect size. Alternatively, we expect the activeness of control group (i.e., a measure of whether control group members sought alternative services on their own) to be negatively associated with average effect size.

Research Methods:

Meta-analysis. To understand how specific features of research design may account for the heterogeneity in estimated Head Start effects, 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, study samples, etc. 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 Head Start 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).

An important first step in a meta-analysis is to identify all relevant evaluations that meet one's programmatic and methodological criteria for inclusion; therefore, a number of search strategies were used to locate as many published and unpublished Head Start evaluations conducted between 1965 and 2007 as possible.[†] First, we conducted key word searches in ERIC, PsychINFO, and Dissertation Abstracts databases. Next, the research team tracked down additional reports mentioned in collected studies. Over 250 new Head Start evaluations were identified, in addition to the 98 coded by Abt. To be exhaustive, the research team will also

[†] The original Abt database included ECE programs evaluated between 1960 and 2003 and used similar search techniques; therefore, we did not re-search for Head Start evaluations conducted during these years, with the exception of 2003. We conducted searches for evaluations completed between 2003 and 2007, as well as for programs not targeted by the original Abt search strategies.

search additional specialized databases, government databases, ECE policy group websites, and conference programs; we will also contact researchers in the field.

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 10 participants in each condition, with attrition of less than 50 percent. 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 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 impact evaluations of Head Start, we impose some additional inclusion criteria. We include only studies that measure differences between Head Start participants and control groups that were assigned to receive no other services. For example, studies that randomly assigned children to Head Start versus another type of early education program or Head Start add-on program are excluded. However, studies are not excluded if children assigned to a no alternative treatment control group sought services of their own volition. In addition, we include only studies that provide at least one measure of children's cognitive or achievement outcomes. Thus far, the screening process, based on the above criteria, has resulted in the inclusion of 48 Head Start publications or reports.[‡]

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 10 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 study, the contrast, and the effect size. A "study" is defined as a collection of comparisons in which the treatment groups are drawn from the same pool of subjects. Each study also produces a number of "contrasts," defined as a comparison between one group of children who received Head Start and another group of children who received no other services. Studies may include multiple contrasts; for example, results may be presented using more than one analytic method (e.g., OLS and fixed effects), and these are coded as different contrasts nested within one study. The 20 Head Start studies currently coded and in the database include 62 separate contrasts. In turn, each contrast provides a number of individual "effect sizes" (estimated standard deviation unit difference in an outcome between the children who experienced Head Start and those who did not). The 62 contrasts in the database provide a total of 377 effect sizes.[§] These effect sizes combine information from a total of 10,268 observations. (See Table 1: Key Meta-Analysis Terms and Ns; please insert Table 1 here).

[‡] 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.

[§] 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.

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, 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.

Fifty contrasts provided more than one measure of cognitive skills or achievement. In these cases, including all effect sizes as separate observations would violate the assumption of statistical independence. We follow standard meta-analysis procedures by aggregating the effect sizes within each contrast (Lipsey & Wilson, 2001). Because a clear "best measure" within each contrast was not readily apparent, we average together all cognitive and achievement effect sizes within a contrast to create one "average effect size" per contrast. Thus, there are currently 62 average effect sizes available for analysis.

Measures. The dependent variables in these analyses are the effect sizes measuring the impact of Head Start on children's cognitive skills and achievement. 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. Currently coded effect sizes have an unweighted mean of .13 and interquartile range of .44.

The key independent variables represent facets of the contrast design. These include type and rigor of design, quality of dependent measure, timing of outcome measure, attrition, and activity level of control group. Specific information explaining how some of these concepts are coded is provided in Table 2: Descriptive Information for a Select Sub-set of Methodological Variables of Interest (please insert Table 2 here).

Although Head Start is guided by a set of federal performance standards and other regulations, these have changed over time, and may not reflect the experience of participants in all studies. Therefore, reports were also coded along dimensions on which programs are expected to vary, such as population served, staff credentials, and dosage. These varying program characteristics may also be used as covariates in analyses. Other relevant covariates may include type of publication, year published, and baseline characteristics of the sample.

Statistical analysis. Our key research question is whether heterogeneity in the average effect size is predicted by methodological aspects of the contrasts. The nested structure of the data (contrasts nested within studies) requires a multivariate, multi-level approach to modeling these associations. The level-1 model (contrast level) is:

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

In this equation, the average effect size (ES_{ij}), for study *i* and contrast *j*, is modeled as a function of the intercept (β_{0i}), which represents the average (covariate adjusted) effect size for all contrasts, a series of key independent variables and related coefficients of interest ($\beta_{1i}x_{1ij} + \ldots + \beta_{ki}x_{kij}$), which estimate the association between the average effect size and coded aspects of the study design as described above, as well as other relevant covariates, and a within-study error term (e_{ij}). The level-2 equation (study level) models the intercept as a function of the grand mean effect size (β_0) and a between-study random error term (u_i):

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. These effect sizes (N=51) may be imputed using multiple imputation techniques (the *ice* program in STATA) for use in our final analyses (Royston, 2004).

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

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-contrast variables that measure key features of the contrast methods and other covariates; and 2) remaining "random" unmeasured sources of variation between and within studies.

We will also test several variations of the model specification described above; for example, we will conduct separate analyses based on the timing of outcomes (separating measures taken at program completion from longer-term follow-up measures) and for policy-relevant population differences, such as three- versus four-year-olds. Given prior research suggesting that some skills are more sensitive to instruction than others (Christian, Morrison, Frazier, & Massetti, 2000), we will also consider separate analyses of achievement outcomes (e.g., measures of early reading and math skills) and other cognitive outcomes (e.g., IQ and vocabulary). Similar to prior meta-analyses, we may also derive composite measures of dependent measure quality (e.g., based on type of measure, reliability, and whether the data collector was blinded) and overall study quality (e.g., based on factors such as type and rigor of design, attrition, and equivalence of treatment and control groups). To account for differences in sample sizes across studies, regressions will also be weighted by the inverse variance weight of each effect size (Lipsey & Wilson, 2001).

Results:

Although we are still completing coding for some Head Start studies that will ultimately be included in our final analyses, preliminary examination of the available data reveals interesting variation along methods-related characteristics between contrasts, as reported in Table 2. For example, we see substantial variation in the activity level of the control group; type of research design; whether baseline equivalency was tested, and if so, whether significant differences between groups were detected; and whether other sources of bias were detected by coders. Additional variation in the type and reliability of dependent measures, timing of tests, and attrition was also detected at the effect size level.

Discussion:

The proposed study provides an important contribution to the field of Head Start research, in that it utilizes a unique, new meta-analytic database to explore the role of methodological factors in explaining variation in effect sizes measuring the impact of the program on children's cognitive skills and achievement. This information can be used by researchers and policy makers to become better consumers and designers of Head Start evaluations; thus, facilitating better policy and program development. Preliminary descriptive analyses of the available data suggest sufficient variation in the methods-related variables to estimate these analyses. We are confident that all coding will be completed by the end of the calendar year, and that we will be able to complete a thorough analysis of the data for presentation at the SREE Conference in March.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Borenstein M., Hedges, L., Higgins, J., & Rothstein, H. (2005). Comprehensive Meta-analysis, Version 2. Englewood NJ: Biostat.

Camilli, G., Vargas, S., Ryan, S., & Barnett, W. S. (2008). Meta-analysis of the effects of early education interventions on cognitive and social development. *Teachers College Record, 112*(3).

Christian, K., Morrison, F. J., Frazier, J. A., & 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. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis, 2nd edition*, (pp. 3-17). New York: Russell Sage Foundation.

Currie, J. & Thomas, D. (1995). Does Head Start make a difference? *American Economic Review, 85,* 341-364.

Gorey, K. M. (2001). Early childhood education: A meta-analytic affirmation of the short- and long-term benefits of educational opportunity. *School Psychology Quarterly, 16*(1), 9–30.

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

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

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

McKey, R. H., Condelli, L., Ganson, H., Barrett, B. J., McConkey, C., & Plantz, M. C. (1985). *The impact of Head Start on children, families and communities: Final report of the Head Start Evaluation, Synthesis and Utilization Project.* Washington, D. C.: CSR, Incorporated.

Nathan, R. P. (Ed.) (2007). How should we read the evidence about Head Start?: Three views. *Journal of Policy Analysis and Management, 26(3),* 673-689.

Nelson, G., & Westhues, A., & MacLeod, J. (2003). A meta-analysis of longitudinal research on preschool prevention programs for children. *Prevention and Treatment, 6*, 1–34.

Office of Head Start (2008). *Head Start program fact sheet, FY 2007*. http://www.acf.hhs.gov/programs/ohs/about/fy2008.html. Washington, DC; Administration for Children, Youth and Families.

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

United States General Accounting Office. (1997). *Head Start: Research provides little information on impact of current program.* GAO/HEHS-97-59. Washington, D. C.: U. S. General Accounting Office.

Zigler, E. & Valentine, J. (Eds.) (1979). *Project Head Start: A legacy of the war on poverty.* New York: The Free Press.

Appendix B. Tables and Figures

Not included in page count.

Table 1: Key Meta-Analysis Terms and Ns

Term	Description	N in current database*
Report	Written evaluation of Head Start (e.g., a journal article, government report, book chapter)	48
Study	Collection of comparisons in which the treatment groups are drawn from the same pool of subjects	20
Contrast	Comparison between one group of children who received Head Start and another group of children who received no other services	62
Effect Size	Measure of the difference in cognitive outcomes between the children who experienced Head Start and those who did not, expressed in standard deviation units (*Hedges' g*)	377
Average Effect Size	When individual contrasts include more than one cognitive effect size, these effect sizes are averaged together to create an "average effect size" aggregated at the contrast level	62

*Note: We estimate that our database currently contains approximately 75 percent of the studies that we will use in the final analyses.

Table 2: Descriptive Information for Selected Sub-set of Methodological Variables of Interest

Contrast Level (N=62)	
Construct	**Description (frequencies)**
Activity level of control group	Active (11) Passive (48) Missing (3)
Type of research design:	Fixed effects, family or individual (22) Change analysis other longitudinal change (3) Matching on demographics or baseline outcome (22) Above quasi-experimental designs don't apply, but baseline equivalent (6) Randomized controlled trial (7)
Baseline equivalence between treatment and control group	Baseline equivalence tested, no significant differences (14) Baseline equivalence tested, significant differences (29) Baseline equivalence not tested (19) (Information is also available regarding which characteristics are significantly different; e.g., pre-test measure, family composition, family SES/education, parenting skills/attitudes, race/ethnicity, gender, child functioning)
Bias	Any additional source of bias identified by coders (24) No additional source of bias identified (38) (Information is also available regarding specific sources of bias; e.g., whether only treated subjects are included in the analysis, or whether the degree of volunteering is different for the treatment and control groups)

Table 2: Descriptive Information for Selected Sub-set of Methodological Variables of Interest, Continued

Effect Size Level (N=377)	
Construct	**Description (frequencies)**
Type of dependent measure	Rating by someone else (19) Performance test (300) Observational rating (25) Missing (33)
Reliability of dependent measure	Reliability data on dependent measure available (114) No reliability data available (263) (Information is also available regarding the type and magnitude of reliability coefficient, and whether the data collector was blinded)
Timing of outcome measure	End of treatment (100) Observation point during treatment (9) Follow-up test after treatment (227) Combination of categories (41) (Information is also available regarding the number of months between beginning of treatment and outcome measure)
Percent attrition*	Minimum=0% Maximum=48% Unweighted mean=13% Missing (12) (Information is also available regarding whether attrition bias was tested, and if so, whether there are significant differences resulting from attrition and type of attrition correction)

*Note: Summary statistics for percent attrition are based on information for 347 effect sizes; 30 observations are not included because additional information needs to be coded in order to accurately calculate the attrition rate.