Educational Evaluation and Policy Analysis
March 2013, Vol. 35, No. 1, pp. 76–95
DOI: 10.3102/0162373712462453

Can Research Design Explain Variation in Head Start Research Results? A Meta-Analysis of Cognitive and Achievement Outcomes

Hilary M. Shager
University of Wisconsin-Madison

Holly S. Schindler
University of Washington, Seattle

Katherine A. Magnuson
University of Wisconsin-Madison

Greg J. Duncan
University of California, Irvine

Hirokazu Yoshikawa
Harvard University

Cassandra M. D. Hart
University of California, Davis

This study explores the extent to which differences in research design explain variation in Head Start program impacts. We employ meta-analytic techniques to predict effect sizes for cognitive and achievement outcomes as a function of the type and rigor of research design, quality and type of outcome measure, activity level of control group, and attrition. Across program evaluations, the average program-level effect size was 0.27 standard deviations. About 41% of the variation in estimates across evaluations can be explained by research design features, including the extent to which the control group experienced other forms of early care or education, and 11% of the variation within programs can be explained by the quality and type of the outcome measures.

Keywords: *meta-analysis, Head Start, program evaluation*

The recognition that school-entry academic skills of poor children lag well behind those of their more advantaged peers has focused attention on early childhood education (ECE) as a potential vehicle for remediating early achievement gaps. Sharp increases in public spending on a variety of ECE programs over the past 20 years reflect the success of educators and advocates in arguing for the value of early education programs for disadvantaged children. This increased attention and funding has also produced a proliferation of ECE program evaluations.

These studies can yield important information about differences in the effectiveness of particular program models, but only if we understand the context of ECE research and are confident that divergent findings reflect meaningful differences in program effectiveness rather than methodological differences in study design.

Researchers and policymakers face a daunting task in making sense of findings across studies. Evaluations vary greatly in methods, quality, and results, which leads to an "apples versus oranges" comparison problem. Previous reviews of ECE programs have described differences in research design as a part of their subjective, narrative analyses and have suggested that such features might be important (Magnuson & Shager, 2010; Ludwig & Phillips, 2007; McKey et al., 1985). However, there has been little systematic empirical investigation of the role of study design features in explaining differing results. This meta-analysis first estimates the short-term cognitive and achievement impacts of Head Start, a federally funded early education program for low-income children, using the accumulated evidence of program evaluations conducted between 1965 and 2007. Second, this study investigates the role of research design in predicting program impacts to better understand the program's effects. Taken together, these analyses will inform readers about the broad set of findings from rigorous Head Start evaluations and will provide important context for interpreting the findings of Head Start and possibly other ECE program evaluations by identifying research design features that are likely to predict larger (or smaller) program effects.

Background

Within the field of ECE, Head Start is the largest publicly funded early education program, serving more than 900,000 children with annual federal funding of more than $7.2 million (U.S. Department of Health and Human Services [DHHS], 2010b). A centerpiece of President Lyndon B. Johnson's "War on Poverty," 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). Despite the program's holistic approach, most Head Start evaluations have focused primarily on academic and cognitive outcomes, with few studies including good measures of social, emotional, or health outcomes.

The extent to which Head Start children benefit from participating in the program remains a topic of debate, which began with the first national Head Start study (Westinghouse Learning Corporation, 1969) and continued following the recent releases of findings from the program's national random-assignment evaluation (U.S. DHHS, 2010, January). The recent findings have been characterized by some critics as showing that Head Start has small and transitory effects on cognitive and achievement outcomes, which they have contrasted with the larger effects of other early education programs. Besharov and Higney (2007) wrote,

> It seems reasonable to compare Head Start's impact to that of both pre-K programs and center-based child care. According to recent studies, Head Start's immediate or short-term impacts on a host of developmental indicators seem not nearly as large as those of either pre-K or center-based child care. (pp. 686–687)

Yet, other scholars have argued that methodological differences may at least in part explain the differences in program impacts across such ECE studies, suggesting that differences in study design make the evaluation results incomparable (Cook & Wong, 2007). How large a role these design features might play, however, has been only a matter of conjecture.

That an evaluation's research design affects study results is widely accepted, and, as a result, experimental methods have been anointed as the gold standard for program evaluation (Cook, Shadish, & Wong, 2008). However, the magnitude and direction of bias from the use of nonexperimental methods are often unknown and may be specific to the program and population of study. Several meta-analyses of medical and social service interventions have found that low-quality study designs yield larger effect sizes than do high-quality ones (Moher et al.,

1998; Schulz, Chalmers, Hayes, & Altman, 1995; Sweet & Appelbaum, 2004). Alternatively, within Head Start research, the primary concern has been that low-quality, nonexperimental studies will understate program effects because they are likely to result in more advantaged comparison groups (Currie & Thomas, 1995).

Within-study comparisons of research designs provide the best way to assess the importance of research design, but such studies are rare. Cook and colleagues' (2008) recent examination of a small number of within-study comparisons of research designs concluded that high-quality nonexperimental study designs (regression discontinuity, matched intact local control groups, and modeling selection processes) can closely approximate experimental studies (also see Shadish, Clark, & Steiner, 2008; Shadish, Galindo, Wong, Steiner, & Cook, 2011). Unfortunately, to date, within-study research design comparisons have not been conducted within the ECE field; thus, although studies from other fields provide important insight into the general role of research methods, it is unclear if these findings will likely apply to Head Start evaluations. Moreover, such studies have not examined other important facets of study design that are likely to affect ECE studies, such as the activity level of the control group and attrition.

Some prior meta-analytic studies of ECE programs have given at least cursory attention to the question of whether study design and related threats to internal validity predict the magnitude of program impacts. The only existing meta-analysis of Head Start research, conducted more than 25 years ago by McKey and colleagues (1985), found short-term positive program impacts on cognitive test scores (effect sizes = 0.31 to 0.59). Initial descriptive analyses of methodological factors related to issues of internal validity, such as quality of study design and attrition, revealed only slight influences on the magnitude and direction of effect sizes, which led the authors to exclude these measures from the main analyses.

Typically, other ECE meta-analytic studies have used composite measures of study design, combining aspects of internal validity with other study features, and it is not surprising that these idiosyncratic researcher-designed measures have yielded mixed findings. Null associations between effect sizes and study design quality composites, which included a variety of indicators related to internal validity, were found in Gorey's (2001) and Nelson, Westhues, and MacLeod's (2003) meta-analyses of ECE programs with long-term follow-up studies. In contrast, Camilli, Vargas, Ryan, and Barnett's (2010) ECE meta-analysis found that studies with a higher quality design yielded larger effect sizes for cognitive outcomes. The measure of design quality was a dichotomous indicator based on such factors as attrition, baseline equivalence of treatment and control groups, high implementation fidelity, and adequate information provided for coders.

In this study, we investigate the importance of several aspects of study design related to internal validity, but do so without aggregating features of study quality into single composites that may obscure the importance of any specific evaluation feature. Considering the role of research methods in evaluation report findings is possible because Head Start has provided a fairly standardized set of services to a relatively homogenous population of children over a long period of time, and numerous evaluations have been conducted. Given the importance of ensuring the baseline equivalence of the treatment and control groups (Currie & Thomas, 1995; Gibbs, Ludwig, & Miller, 2012), *we hypothesize that studies that used more rigorous methods to ensure similarity between treatment and control groups prior to program participation, particularly random assignment, will produce larger effect sizes*. Likewise, since disadvantaged students most likely to benefit from the program are also most likely to be lost to follow-up, *we predict that studies with higher levels of overall attrition may yield smaller effect sizes*.

Another feature of ECE evaluation design overlooked in prior studies is the activity level of the control group. In the case of ECE evaluations, this is defined as participation in center-based care or preschool among control group children. Comparison group children's participation in other early education and care programs is important because low-income children gain important academic and cognitive benefits from attending these programs. For example, Zhai,

Brooks-Gunn, and Waldfogel (2010) found that when Head Start attendees were compared with children who received parental or other informal care, they had better academic outcomes; however, they did no better on these outcomes when compared with children who attended other center-based care programs.

Control group activity varies considerably across ECE studies and is particularly high in more recent evaluations, given the high rates of participation in early care and education programs among children of working parents and the expansion of ECE programs in the decades since Head Start began. The issue of the activity of the control group figured prominently in discussions about the recent national Head Start evaluation (National Forum on Early Childhood Policy and Programs, 2010). Cook (2006) found that rates of participation of the control group in ECE settings during the preschool "treatment" year were more than 50% in the Head Start Impact Study (U.S. DHHS, Administration for Children and Families, 2005), considerably higher than the approximately 20% found in state prekindergarten studies (Wong, Cook, Barnett, & Jung, 2008). He argued that these differences might in part explain the larger effect sizes associated with recent prekindergarten evaluations compared with the findings reported in the Head Start Impact Study. Our study is the first to empirically investigate whether the activity level in the control group predicts the magnitude of program impact effect sizes across Head Start evaluations. *We hypothesize a negative relationship between effect size and the activity level of the control group* (i.e., a measure of whether control group members sought alternative ECE services on their own).

Other features of study design also may predict evaluations' effect sizes—in particular, the characteristics of the measures selected. First, the alignment between the intervention content and the skills being assessed may matter because academic skills are more likely to be improved by ECE and other instructional settings than general cognitive skills (Christian, Morrison, Frazier, & Massetti, 2000). Suggestive evidence that measures that are more closely aligned with the practices of classroom instruction may be more sensitive to early education is found in Wong and colleagues' (2008) regression discontinuity study of five state prekindergarten programs. All of the programs measured both children's receptive vocabulary and their level of print awareness. Across the five programs, effects on print awareness were several times greater than the effects on receptive vocabulary. *We hypothesize that types of skills that are more closely aligned with early education instruction, such as prereading and premath academic skills, will yield larger effect sizes than measures of more abstract and global cognitive skills*, *such as vocabulary and IQ*.

A second feature related to measurement, reliability, has also been overlooked in prior studies. In general, measurement error is likely to attenuate associations between variables; thus, we might expect larger estimates from more reliable measures. However, if low reliability is a proxy for evaluator-developed tests, which indicates the assessment is likely to be closely aligned with the skills taught in a particular Head Start program, then low reliability may predict larger program impact effect sizes compared with standardized tests (Rosenshine, 2010). *Given this uncertainty, we do not have a clear hypothesis about the direction and magnitude of the association between measure reliability and Head Start program impact effect sizes and consider our analysis exploratory.*

A third feature of measures is the method of assessment, specifically whether data are collected via a direct assessment on a particular test or task or as an observational rating. Standardized direct assessments, which are designed to have high levels of reliability, are likely to introduce less measurement error than either teacher or parent reports of children's cognitive skills. Hoyt and Kerns's (1999) meta-analysis of psychological studies found that rating bias was likely to be very low in measures that included explicit attributes (counts of particular behaviors) but quite prevalent in measures that required raters to make inferences (global ratings of achievement or skills). If ratings of young children elicit global ratings, then such assessments might be more biased than direct assessments of children's skills. However, such a prediction is complicated by the fact that observer reports may be more aligned with skills taught in particular programs, which would suggest that these assessments may predict larger effect sizes. Moreover, it is likely

that in many studies, raters, especially teachers and parents, were not blind to the children's treatment status, in which case it is possible that the resulting bias would favor Head Start attendees and predict larger effect sizes (Hoyt & Kerns, 1999). *Given competing arguments, we consider our examination of the method of assessment (direct assessment versus observation) to be exploratory and do not have a clear hypothesis about the direction or magnitude of its association with effect sizes.*

Research Method

Meta-Analysis

To summarize more than 30 years of Head Start research and understand how specific features of research design may account for the heterogeneity in estimated Head Start effects, we conducted 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, an effect size, which expresses treatment-control differences as a fraction of the standard deviation of the given outcome. Outcomes from individual studies can then be used to estimate the average effect size across studies. In addition, meta-analysis can be used to test whether the average effect size differs by, for example, type and quality of research design. After defining the problem of interest, meta-analysis proceeds in the following steps, described below: (a) literature search, (b) data evaluation, and (c) data analysis.

Literature Search

The Head Start studies analyzed in this article compose a subset of studies from a large meta-analytic database being compiled by the National Forum on Early Childhood Policy and Programs. 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 previous meta-analytic databases created by Abt Associates and the National Institute for Early Education Research (NIEER; Camilli et al., 2010; 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 keyword searches in the ERIC, PsycINFO, EconLit, and Dissertation Abstracts databases, resulting in 304 Head Start evaluations. Next, we manually searched the websites of several policy institutes (e.g., RAND, Mathematica, NIEER) and state and federal departments (e.g., U.S. DHHS) as well as references mentioned in collected studies and other reviews. This search resulted in another 134 possible reports for inclusion in the database. In sum, 438 Head Start evaluations were identified, in addition to the 126 previously coded by Abt and NIEER.

Data Evaluation

The next step in the meta-analysis process was to determine whether identified studies met our established inclusion criteria. To be included in our database, studies must have had (a) a comparison group (either an observed control or alternative treatment group) and (b) at least 10 participants in each condition, with attrition of less than 50% in each condition. Evaluations could have been experimental or quasi-experimental, using one of the following methods: regression discontinuity, fixed effects (individual or family), residualized or other longitudinal change models, difference in difference, instrumental variables, propensity score matching, or interrupted time series. Quasi-experimental evaluations not using one of the former analytic strategies were also screened in if they included a comparison group *plus* pre- and posttest information on the outcome of interest or demonstrated adequate comparability of groups on baseline characteristics. These criteria are more rigorous than those applied by McKey et al. (1985), Abt, and NIEER; for example, they eliminated all pre/post-only (no comparison group) studies as well as regression-based studies with baseline nonequivalent treatment and control groups.

For this particular study, we imposed some additional inclusion criteria. We included only studies that measure differences between Head Start participants and control groups that were assigned to receive no other services. For example, studies that compared Head Start attendees with children who attended another type of early education program or Head Start add-on program were excluded. However, studies were not excluded if families assigned to a no-treatment control group sought services of their own volition. In addition, we included only Head Start studies that had at least one measure of children's cognitive or achievement outcomes. Outcome measures from other domains, such as socioemotional development and health, were too rare to be analyzed separately. Furthermore, to improve comparability across findings, we imposed limitations regarding the timing of study outcome measures. We limited our analysis to studies in which children received at least 75% of the intended Head Start treatment and for which outcomes were measured 12 or fewer months posttreatment.

The screening process, based on the above criteria, resulted in the inclusion of 57 Head Start publications or reports (see the appendix). Of the 126 Head Start reports originally included in the Abt/NIEER database, 29 were eliminated from the database because they did not meet our research design criteria. The majority of the 438 additional reports identified by the research team's search were excluded after reading the abstract (n = 243), indicating that they did not meet our inclusion criteria for obvious reasons (e.g., they were not quantitative evaluations of Head Start or did not have a comparison group). Of the 98 Head Start evaluation reports that were excluded after full-text screening, 90 were excluded because they did not meet our research design criteria; 53 of these were excluded specifically because they did not include a comparison group. Eight other reports were excluded based on other eligibility criteria (e.g., they reported results only for students with disabilities or did not include relevant outcomes). Our additional inclusion criteria for this article (e.g., short-term cognitive or achievement outcomes, no alternative treatment or curricular add-on studies) excluded 120 reports that remain coded in the database but are not included in this analysis.

Coding Studies

The research team developed a protocol to codify information about study design, program and sample characteristics, and 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. Trainees were paired with experienced coders in multiple rounds of practice coding. Before coding independently, research assistants passed a reliability test comprising randomly selected codes from a randomly selected study. To pass the reliability test, researchers had to calculate 100% of the effect sizes correctly and achieve 80% reliability with a master coder for the remaining codes. In instances when research assistants were just under the threshold for effect sizes but were reliable on the remaining codes, they underwent additional effect size training before coding independently and were subject to periodic checks during their transition. Questions about coding were resolved in weekly research team conference calls.

Database

The resulting database is organized in a four-level hierarchy (from highest to lowest): the study, the program, the contrast, and the effect size (see Table 1). A "study" is defined as a collection of comparisons in which the treatment groups are drawn from the same pool of participants. A study may include evaluations of multiple "programs;" i.e., a particular type of Head Start or Head Start in a particular location. 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 as a result of the study. Evaluations of programs within studies may include multiple con-

TABLE 1
Key Meta-Analysis Terms and n *Values*

Term	Description	*n*
Report	Written evaluation of Head Start (e.g., a journal article, government report, book chapter)	57
Study	Collection of comparisons in which the treatment groups are drawn from the same pool of participants	27
Program	Particular type of Head Start or Head Start within a particular location	33
Contrast	Comparison between one group of children who received Head Start and another group of children who received no other services as a result of the study (although they may have sought services themselves)	40
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's *g*)	313

trasts; for example, results may be presented using more than one analytic method (e.g., ordinary least squares and fixed effects) or separate groups of children (e.g., 3- and 4-year-olds), and these are coded as different contrasts nested within one program, within one study. We include 33 Head Start programs in our analysis, but 5 of these programs provided only "missing" effect sizes (insufficient detail was provided in the report to calculate an effect size), and, as a result, our primary analyses consist of data from 28 programs. The 33 Head Start programs included in our meta-analysis include 40 separate contrasts of program main effects (excluding subgroup analyses, e.g., by gender or race). In turn, each contrast consists of 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 40 contrasts in this meta-analytic database provide a total of 313 effect sizes (72 of these effect sizes were coded as "missing"). These effect sizes combine information from a total of more than 160,000 observations. See the appendix for a description of the Head Start studies, programs, and contrasts included in our analyses.

Effect Size Computation

Outcome information was reported in evaluations using a number of different statistics, which were converted to effect sizes (Hedges's *g*) with the commercially available software package Comprehensive Meta-Analysis (Borenstein, Hedges, Higgins, & Rothstein, 2005). Hedges's *g* is an effect size statistic that adjusts the standardized mean difference (Cohen's *d*) to account for bias in the *d* estimator from small sample sizes.

Dependent Variable

Descriptive information for the dependent measure (effect size) and key independent variables is provided in Table 2. To account for the varying precision among effect size estimates, as well as the number of effect sizes within each program, these descriptive statistics and all subsequent analyses are weighted by the inverse of the variance of each effect size multiplied by the inverse of the number of effect sizes per program (Cooper & Hedges, 2009; Lipsey & Wilson, 2001). The dependent variables in these analyses are the effect sizes measuring the standardized difference in assessment of children's cognitive skills and achievement between children who attended Head Start and the comparison group. Effect sizes ranged from –0.49 to 1.05, with an estimated weighted mean of 0.18 at the effect size level and an estimated weighted mean of 0.29 at the program level.

Independent Variables

Several key independent variables represent facets of each program's research design. These study and Head Start characteristics do not vary within a program, so we present both program-level and effect-size-level descriptive information for these measures in Table 2. We created a dummy variable indicating whether a study was randomized (reference category) or quasi-experimental. Two programs had designs that

TABLE 2
*Descriptive Information for Nonmissing Effect Sizes and Independent Variables (*N = *241)*

	Min	Max	*M*	*SD*
Study and program characteristics				
Modern Head Start program (post-1974), effect size level	0	1	0.44	0.50
Modern Head Start program (post-1974), program level	0	1	0.25	0.44
Length of program (months, centered at 2), effect size level	0	8	6.20	3.12
Length of program (months, centered at 2), program level	0	8	5.18	3.70
Peer-refereed journal, effect size level	0	1	0.17	0.37
Peer-refereed journal, program level	0	1	0.21	0.42
Design characteristics				
Active control group, effect size level	0	1	0.35	0.48
Active control group, program level	0	1	0.14	0.36
Passive control group, effect size level	0	1	0.53	0.50
Passive control group, program level	0	1	0.68	0.48
Missing control group activity, effect size level	0	1	0.12	0.32
Missing control group activity, program level	0	1	0.18	0.39
Randomized controlled trial, effect size level	0	1	0.26	0.44
Randomized controlled trial, program level	0	1	0.18	0.39
Quasi-experimental study, effect size level	0	1	0.74	0.44
Quasi-experimental study, program level	0	1	0.82	0.39
Baseline covariates included, effect size level	0	1	0.49	0.50
Baseline covariates included, program level	0	1	0.14	0.36
Dependent measure characteristics				
Rating (by someone who knows child)	0	1	0.06	0.24
Observation (by researcher)	0	1	0.01	0.09
Performance measure	0	1	0.93	0.25
Skills not sensitive to instruction	0	1	0.67	0.47
Skills sensitive to instruction	0	1	0.33	0.47
Months posttreatment	–2.47	12	1.89	4.12
Attrition (always < 50%)				
High attrition (> 10%)	0	1	0.36	0.48
Low attrition (≤ 10%)	0	1	0.42	0.50
Missing attrition information	0	1	0.22	0.41
Reliability				
High reliability (coefficient ≥ .92)	0	1	0.21	0.41
Medium reliability (coefficient = .75–.91)	0	1	0.35	0.48
Low reliability (coefficient < .75)	0	1	0.18	0.39
Missing reliability coefficient	0	1	0.26	0.44
Effect size, effect size level	–0.49	1.05	0.18	0.22
Effect size, program level	–0.23	0.78	0.29	0.23

Note. Weighted by inverse variance of effect size multiplied by inverse of number of effect sizes per program.

changed post hoc (the study was originally randomized but, for various reasons, became quasi-experimental in nature). In our primary specifications, these programs were coded as quasi-experimental. The majority of effect sizes (74%) come from quasi-experimental studies, although differences between program level and effect size level means suggest that randomized trials tended to have more outcome measures per study than those with quasi-experimental designs. In addition, we created a dummy variable to indicate whether baseline covariates were included in the analysis. Although the majority of programs (86%) did not include baseline covariates in their analyses, those that did had a large number of outcome measures.

We created a series of dummy variables indicating levels of overall attrition. Keeping in mind that attrition was truncated at 50% based on our screening criteria, attrition levels were constructed using quartile scores and defined as follows: low attrition (reference category), less than or equal to 10% (representing Quartiles 1, 2, and 3); high attrition, greater than 10% (repre-

senting Quartile 4); and missing overall attrition information. The plurality of effect sizes came from studies with 10% attrition or less.

We also coded the activity level of the control group using the following categories: passive (reference category), meaning that control group children received no alternative services; active, meaning some of the control group members sought services of their own volition; and a dummy variable indicating whether information regarding control group activity was missing from the report. Reports in which there was no mention of control group activity were coded as having missing information on this variable. Although the majority of effect sizes (53%) came from studies with passive control groups, studies in which the control group actively sought alternative services, specifically attendance at other center-based child care facilities or early education programs, tended to have more effect sizes per study. Studies that reported active control groups indicated that between 18% and 48% of the control group attended these types of programs.

Several of the key independent variables describe features of the outcome measures. We distinguished between effect sizes measuring achievement outcomes, such as reading, math, letter recognition, and numeracy skills, which may be more sensitive to typical classroom instruction, and those measuring cognitive outcomes less sensitive to instruction, including IQ, vocabulary, theory of mind, attention, task persistence, and syllabic segmentation, such as rhyming (see Christian et al., 2000, for a discussion of this distinction). The majority of effect sizes (67%) were from the cognitive domain. Using a series of dummy variables, we also categorized effect sizes according to the type of measure employed by the researcher, indicating whether it was a performance test (reference category), a rating by someone the child knows (e.g., a teacher or parent), or an observational rating by a researcher. The majority of outcome measures were performance tests (93%).

Control Variables

By limiting this study to Head Start evaluations, our analyses hold constant program features such as funding stream, program structure and requirements, and family socioeconomic background of children served. We do include other measured features of the evaluation studies in our analyses as controls because they may be confounded with research study design. 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. A dummy variable was coded to indicate whether the program was a "modern" Head Start program, defined as post-1974, when the first set of Head Start quality guidelines were implemented. Although the majority of programs (75%) were older, 44% of effect sizes came from studies of modern Head Start programs.

Recognizing that the first iteration of Head Start was a shortened 6- to 8-week summer program, we also created a continuous variable indicating length of treatment measured in months and recentered at 2 months, so that the resulting coefficient indicates the effect of receiving a full academic year of Head Start versus a summer program.

Evaluations differ in the timing of the outcome assessments, and given that Head Start program impacts are found to decline over time, we included a continuous measure of the timing of the outcome, measured in months posttreatment. Given our screening criteria, this variable ranges from –2.47 to 12. Finally, we created a dummy variable indicating whether the evaluation was an article published in a peer-refereed journal. The reference category is an unpublished report, dissertation, or book chapter.

Statistical Analysis

Our key research questions are the following: (a) What is the average program impact of Head Start on children's cognitive and achievement outcomes? and (b) Is heterogeneity in effect sizes predicted by methodological aspects of the study design and attributes of the outcome measures? The nested structure of the data (effect sizes nested within programs) requires a multivariate, multilevel approach to modeling these associations (de la Torre, Camilli, Vargas, & Vernon, 2007). The Level 1 model (effect size level) is:

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

In this equation, the effect size *j* in program *i* is modeled as a function of the intercept (β_{0i}), which represents the average (covariate adjusted) effect size for all programs; 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 effect size and aspects of the study design that vary at the effect size level; and a within-program error term (e_{ij}). Study design covariates at this level include timing of outcome, type of outcome (rating or observation), whether or not baseline covariates are included, and domain of outcome (skills more or less sensitive to instruction).

The Level 2 equation (program level) models the intercept as a function of the grand mean effect size (β_{00}), a series of covariates that represent aspects of study design and Head Start features that vary only at the program level ($\beta_{01i}x_{1i} + \ldots + \beta_{0ki}x_{ki}$), and a between-program random error term (u_i):

$$\beta_{0i} = \beta_{00} + \beta_{01i}x_{1i} + \ldots + \beta_{0ki}x_{ki} + u_i \quad (2)$$

Variables at this level include type of research design, activity level of control group, and attrition. Covariates at this level include whether the effect size came from a peer-refereed journal article, the length of program, and whether the program was implemented post-1974.

This "mixed effects" model assumes that there are two sources of variation in the effect size distribution, beyond subject-level sampling error: (a) the "fixed" effects of variables that measure key features of the methods and other covariates and (b) remaining "random" unmeasured sources of variation between and within programs. Because of the small number of studies with multiple contrasts ($n = 8$) and multiple programs ($n = 4$), we do not model the third or fourth levels of nesting in our data (contrasts within programs and programs within studies, respectively).

To account for differences in effect size estimate precision, as well as the number of effect sizes within a particular program, all regressions were weighted by the inverse variance of each effect size multiplied by the inverse of the number of effect sizes per program (Cooper & Hedges, 2009; Lipsey & Wilson, 2001). Analyses were conducted in SAS, using the PROC MIXED procedure.

We began by entering each design factor independently and then included all relevant design covariates at the same time in our primary specification. We also tested several variations of the primary model specification; for example, we tested alternative specifications using a series of dummy variables indicating outcome measure reliability levels and a more nuanced set of research design variables. We conducted analyses including imputed missing effect sizes, without weights, and excluding the National Head Start Impact Study, the largest study in our sample.

Results

Bivariate Results

The weighted result from an "empty model," with no predictor variables, yields an intercept (average program-level effect size) of 0.27, which is significantly different from 0. As would be expected, this is very similar to the weighted mean estimated at the program level (0.29) but larger than the weighted mean at the effect size level (0.18). This illustrates the fact that there is more variation within evaluation programs than between programs (in both the weighted and unweighted data). In the weighted data, 80% of the variation in effect sizes is found within programs (and only 20% between programs). Perhaps somewhat surprisingly, programs that had negative or small effect sizes also tended to have large positive effect sizes.

Next, we turned to estimating the associations between single design factors and average effect size using a series of multilevel regressions (Table 3). Regressions including categorical variables (Table 3, columns 1–6 and 8–9) were run without intercepts; thus, the resulting coefficients indicate the average weighted effect size for programs in each category. Multilevel regressions including continuous measures of research design were run with an intercept; therefore, we include this estimate in columns 10 and 11 of Table 3 to show the relationship

TABLE 3

Summary of Results From Regressions of Head Start Evaluation Effect Sizes on Single Research Design Factors

	1	2	3	4	5	6	7—Significant differences
Modern HS (post-1974)	.23*						
	(.08)						
Not modern HS (pre-1975)	.29**						
	(.05)						
Active control group		.08					Passive†
		(.10)					
Passive control group		.31**					Active†
		(.05)					
Missing control group activity		.29*					
		(.10)					
Randomized controlled trial			.33*				
			(.10)				
Quasi-experimental study			.26**				
			(.05)				
Rating (by adult who knows child)				.45**			Performance**
				(.07)			
Observation (by researcher)				.55**			Performance†
				(.16)			
Performance test				.24**			Rating**, observation†
				(.04)			
Skills sensitive to instruction					.40**		Skills not sensitive**
					(.05)		
Skills not as sensitive to instruction					.25**		Skills sensitive**
					(.05)		
High attrition (> 10%)						.28**	
						(.05)	
Low attrition (≤ 10%)						.33**	
						(.06)	
Missing attrition information						.14	
						(.11)	

	8	9	10	11	12—Significant differences
Baseline covariates included	.20*				
	(.03)				
Baseline covariates not included	.29**				
	(.04)				
Peer-refereed journal		.43**			
		(.09)			Not peer-refereed†
Not peer-refereed journal		.23**			
		(.04)			Peer-refereed†
Length of program (months, centered at 2)			.01		
			(.01)		
Months posttreatment				–.00	
				(.01)	
Intercept			.21*	.28**	
			(.07)	(.04)	

Note. Multilevel models were estimated and regression coefficient estimates are reported with standard errors in parentheses provided below the estimates; $N = 241$ effect sizes nested in 28 programs. For columns 1–6 and 8–9, no intercept was estimated; therefore, the resulting coefficients represent the average effect size for programs in each category. Columns 7 and 12 list within-factor categorical means that are statistically significant compared with the indicated category. Multilevel regressions with continuous measures of research design were run with an intercept; therefore, estimates in columns 10 and 11 show the relationship between an incremental increase in each continuous design variable and average effect size.

†$p < .10$. *$p < .05$. **$p < .001$.

between an incremental increase in each continuous design variable and average effect size.

Two features of the evaluation studies that we thought might predict the magnitude of effect sizes did not—analytic design (experimental vs. quasi-experimental and the inclusion of baseline covariates) and level of program or study attrition. One hypothesis was confirmed. We found that evaluation studies in which the control group actively sought alternative ECE services produced a smaller average effect size (0.08) than studies with passive control groups (0.31).

We found features of the outcomes themselves were quite strong predictors of program impact effect sizes. Consistent with our hypothesis, measures of skills more sensitive to instruction yielded a significantly larger average effect size (0.40) than measures of broader cognitive skills (0.25). Finally, we found that ratings (0.45) and observations (0.55) yielded significantly larger effect sizes than performance tests (0.24).

Only one of the covariates predicted effect size magnitudes. The average effect size from a study published in a peer-refereed journal (0.43) was larger than one produced by an unpublished study or book chapter (0.23), but this difference was only marginally significant. Whether the program was "modern," the number of months since program completion, and the length of program were not associated with the size of effects.

Multivariate Results

Our bivariate approach to modeling the predictive power of various study design features ignores the potential important confounds of other design variables; thus, it might yield biased results. Therefore, in our preferred primary specification, we included all design variables at once to investigate the independent and comparative role of each in affecting average effect size (column 1 of Table 4). Coefficients from multivariate models indicate the strength of associations between our independent variables (measuring facets of research design) and effect sizes (differences between treatment and control groups expressed as a fraction of a standard deviation).

In terms of program and study characteristics, we found that most study design features that were statistically significant in our bivariate analyses remained significant in our multivariate analyses. Results indicated a large negative association between effect size and having an active control group in which families independently sought alternative services (–0.33). Likewise, characteristics of the measures again predicted the magnitude of effect sizes. Compared with performance tests, both ratings by teachers and parents, as well as observational ratings by researchers, yielded larger effect sizes (0.16 and 0.32, respectively). As expected, measures of skills more sensitive to instruction produced larger effect sizes than those for less teachable cognitive skills (0.13).

If one considers a performance test a more reliable measure of children's skills, compared to ratings by others, the fact that performance tests predicted smaller effect sizes is somewhat surprising. To test the role of measure reliability more directly, in an alternative specification we removed the variables indicating type of dependent measure (i.e., rating, observation) and instead included a series of dummy variables indicating the level of reliability of the outcome measure, based on coded reliability coefficients.[1] Consistent with our primary specification findings, we found that *less* reliable measures yielded larger effect sizes (see Table 4, column 2).

In terms of the control variables, again, we found evidence of possible publication bias, with studies published in peer-refereed journals also tending to yield effect sizes 0.28 standard deviations larger than those found in unpublished reports, dissertations, or book chapters. In addition, attending a full academic year of Head Start (10 months) emerged as marginally associated with a 0.16 standard deviation unit larger effect than attending a summer Head Start program (2 months).

Robustness Checks

We undertook several additional analyses to determine the sensitivity of our findings to alternative model specifications. Most important, in some cases, authors reported that groups were compared on particular tests but did not report

TABLE 4
Summary of Results From Multivariate Regressions of Head Start Evaluation Effect Sizes on Multiple Research Design Factors

	1	2
Intercept	.30*	.33*
	(.14)	(.15)
Modern Head Start program (post-1974)	–.04	–.16
	(.12)	(.14)
Length of program (months, centered at 2)	.02†	.02†
	(.01)	(.01)
Peer-refereed journal	.28*	.31*
	(.09)	(.11)
Active control group	–.33*	–.35*
	(.12)	(.14)
Missing control group activity	–.01	–.07
	(.10)	(.12)
Quasi-experimental study	–.14	–.26*
	(.11)	(.12)
Baseline covariates included	–.11	–.06
	(.09)	(.10)
High attrition (> 10%)	–.09	–.09
	(.07)	(.07)
Missing attrition information	–.04	.01
	(.13)	(.15)
Rating (by someone who knows child)	.16*	
	(.06)	
Observation (by researcher)	.32*	
	(.15)	
Skills sensitive to instruction	.13**	.17**
	(.03)	(.03)
Months posttreatment	.00	.00
	(.01)	(.01)
Medium reliability (coefficient = .75–.91)		.09†
		(.05)
Low reliability (coefficient < .75)		.20**
		(.06)
Missing reliability coefficient		.07†
		(.04)

Note. Multilevel models were estimated with $N = 241$ effect sizes nested in 28 programs. Regression coefficient estimates are reported with standard errors provided in parentheses below the coefficients.
†$p < .10$. *$p < .05$. **$p < .001$.

the results of these tests, or did not provide enough numerical information to compute an effect size ($n = 72$). We checked the robustness of results to four different assumptions about the magnitude of the missing effects and found that including the imputed effect sizes did not yield substantive changes in our findings (results available from the authors upon request).

We also tested a more nuanced set of research design indicators, distinguishing among types of nonexperimental studies (matching methods, baseline comparable treatment and control groups without matching, use of pretests), and none of these design indicator variables was statistically significant (compared with the reference category, randomization). We also estimated a model with a continuous measure of the year in which the program was studied, and results were similar. Finally, a possible remaining source of heterogeneity in our data is the demographic makeup of the study samples. Unfortunately, only about half of the studies reported demographic characteristics of

the study samples in a way that we could quantify. Nevertheless, with the information available, we explored whether effect sizes might be predicted by the gender and by the racial composition of the sample (percentage boys versus girls; percentage Black, Hispanic, White). Findings indicated that effect sizes were not significantly predicted by these characteristics, in bivariate or multivariate models. Taken together, these findings suggest that our results are robust to a variety of alternative specifications and are not sensitive to the particular models we estimated.

Another concern is that results were being driven primarily by the National Head Start Impact Study, which includes 40 effect sizes and is heavily weighted because of its large sample size. When we excluded this study from our analysis, however, we obtained largely similar results. The magnitude of most coefficients stayed the same, although, because of a loss of statistical power, some of them became statistically insignificant (results available from the authors upon request). These findings suggest that the relationships between effect sizes and research design factors are not unique to the National Head Start Impact Study. Finally, we also found similar results for unweighted analyses, suggesting that studies with larger samples are not driving our findings.

Discussion

This study provides an important contribution to the field of ECE research by using a new meta-analytic database to estimate the overall average effect of Head Start on children's cognitive skills and achievement and exploring the role of methodological factors in explaining effect size variation. Understanding the role of research study design and methods is important so that scholars and policymakers better understand the empirical evidence about Head Start's effectiveness. Overall, we found a statistically significant average effect size of 0.27, suggesting that the accumulated evaluation studies find that Head Start is effective in improving children's short-term (less than 1 year posttreatment) cognitive and achievement outcomes. Several research design factors significantly predicted the heterogeneity of effect sizes. These factors accounted for approximately 41% of the variation between evaluation findings and 11% of the variation within evaluations, suggesting that evaluation research design can be quite consequential.

The resulting average 0.27 effect size suggests that Head Start program effects on children's cognitive and achievement outcomes are on par with the effects of other large-scale ECE programs. This is a somewhat smaller effect on achievement and cognitive outcomes than found in the previous meta-analysis of Head Start conducted by McKey et al. (1985) but larger than those reported in the first-year findings from the recent National Head Start Impact Study (U.S. DHHS, Administration for Children and Families, 2005). The 0.27 estimate is also within the range of the overall average effect sizes on cognitive outcomes found in Camilli et al. (2010), measured across a wider set of ECE programs, and in Wong et al. (2008), measured in state prekindergarten programs, but smaller than the short-term cognitive effect sizes found in meta-analyses of more intensive programs with longitudinal follow-ups conducted (Gorey, 2001; Nelson et al., 2003).

Our substantively largest finding is that having an active control group—one in which children experienced other forms of center-based education and care—is associated with much smaller effect sizes than those produced by studies in which the control group is "passive" (i.e., receives no alternative ECE). Given that other types of ECE and center-based child care programs also increase children's cognitive skills and achievement, this is to be expected (Gormley, Phillips, & Gayer, 2008; Henry, Gordon, & Rickman, 2006). Today, almost 70% of 4-year-olds and 40% of 3-year-olds attend some form of ECE (Cook, 2006); thus, an active control group is likely to be the norm. As a result, Head Start evaluations in communities where many of the control group children have access to other ECE programs are likely to produce substantially smaller effect sizes than those in communities where there are few other ECE programs available. Such a pattern of small or even null effect sizes does not indicate that Head Start is ineffective at improving low-income children's achievement and cognitive outcomes but fol-

lows from the fact that an array of other public and private ECE programs are both accessible and effective in improving low-income children's cognitive and achievement outcomes (Zhai, Waldfogel, & Brooks-Gunn, 2010).

Our analysis is limited to Head Start evaluations, and it is unclear if our findings about the importance of active control groups will generalize to the evaluation of other ECE programs. If these findings were replicated with other types of programs, this result would also suggest that findings from the National Head Start Impact Study, which had relatively high rates of center-care attendance in the control group, and recent regression discontinuity design evaluations of state prekindergarten programs that have lower levels of such attendance in the control group are not directly comparable (Cook, 2006). Thus, claims by some that Head Start is less effective than state prekindergarten programs seem premature, as the higher rate of center-based care and other ECE program attendance among the control group would predict smaller effect sizes for the Head Start study (also see Gibbs, Ludwig, & Miller, 2012).

A somewhat surprising finding from the current study is that the type of overall design (e.g., randomized vs. quasi-experimental) did not predict effect size. Based on prior work, we expected that more rigorous designs would yield larger effect sizes; however, our inclusion criteria regarding study design were typically more rigorous than those of previous meta-analyses of ECE programs. By limiting our study sample in this way, we give up some of the variation in design that might otherwise have predicted effect sizes. Nevertheless, these findings are in alignment with those of recent within-study research suggesting that in certain circumstances rigorous quasi-experimental methods can produce causal estimates similar to those produced by randomized controlled trials (Cook et al., 2008) and further support the use of such methods to evaluate programs when randomized controlled trials are not feasible, as is often the case in education research (Schneider et al., 2007).

We also predicted that attrition would be negatively associated with effect size; however, attrition was not a significant factor. The fact that the range of each measure was truncated in this study (attrition to less than 50% and post-treatment outcome measure timing to 12 or fewer months) may explain this lack of findings. In addition, we were able to investigate only overall study attrition, and it may be that what is most important in predicting program impacts is differential attrition across the control and treatment groups.

For these achievement and cognitive outcomes, more variability was found within program evaluations than across them, suggesting that features of the measures themselves are likely to be an important source of heterogeneity in program evaluation results. We found that the type of dependent measure is systematically related to effect size. Consistent with previous research, we found that achievement-based skills such as early reading, early math, and letter recognition skills appear to be more sensitive to Head Start attendance than cognitive skills such as IQ, vocabulary, and attention, which are less sensitive to classroom instruction (Christian et al., 2000; Wong et al., 2008). This finding has important implications for designers and evaluators of early intervention programs, namely, that expectations for effects on omnibus measures such as vocabulary or IQ should be lowered. At minimum, these sets of skills should be tested and considered separately.

Our finding that less reliable dependent measures yield larger effect sizes also argues for considering the quality of the measures when interpreting program evaluations results. Nonstandardized measures developed by researchers may tap into behaviors that are among those most directly targeted by the intervention services; therefore, it is not surprising that such measures tend to yield larger effect sizes. Ratings by parents, teachers, and researchers may also be subject to bias, however, because these individuals are likely to be aware of children's participation in Head Start as well as the study purpose. In assessing program effectiveness, it is important to compare measures that are similar not only in content but also in method of assessment.

Although they were not the focus of our analyses, our control measures yielded some interesting results. We found that effect sizes from studies published in peer-refereed journals

are larger than those found in unpublished reports and book chapters. Although research published in peer-refereed journals may be more rigorous than that found in unpublished sources, this result may also be a sign of the "file drawer" problem (i.e., that negative or null findings are less likely to be published) long lamented by meta-analysts (Lipsey & Wilson, 2001). This finding suggests that meta-analysts must be exhaustive in their searches for both published and unpublished studies and should code information regarding study quality (Rothstein & Hopewell, 2009).

In addition to the limitations already noted, we offer a few other caveats about this study. First, the nested nature of our data also posed analytic challenges that we could not overcome. Although our multilevel models account for the nesting of effect sizes within programs, there were additional sources of nonindependence in the data set that we were unable to model.

Second, the variation in methods and research design we exploited is naturally occurring; thus, our results are descriptive rather than causal. We modeled characteristics of the assessment measures at the effect size level, so these estimates are identified by variation in measures within evaluations. But our estimates of the predictive power of research design are identified by variation across evaluations and may be biased by unobserved correlates such as the quality of the program or the characteristics of local communities. Studies in other fields have minimized such bias by capitalizing on within-study variation to understand the role of research design, but, to date, such efforts have focused only on understanding under what conditions quasi-experimental designs replicate experimental designs (Cook et al., 2008; Shadish et al., 2008; Shadish et al., 2011). Our meta-analysis is a useful and robust method to summarize studies from Head Start and to improve our understanding of how research design predicts effect sizes. Nevertheless, this study should serve only as an important first step in this line of inquiry. Future studies should undertake within-study design comparisons such as those reviewed in Cook and colleagues (2008), as these have the advantage of controlling for unobserved differences across evaluations.

By analyzing the findings of 28 Head Start evaluations, this study makes an important contribution to the field of ECE research. Although a substantial body of literature indicates that Head Start has meaningful effects on children's cognitive and achievement outcomes, there are a variety of factors that might explain the magnitude of Head Start effects that are about the evaluation research designs, not just the program itself. Thus, comparing results from evaluation studies requires attention to both how the studies are conducted and which instruments are used to assess outcomes. By becoming more critical consumers and designers of such research, we can better understand how well Head Start programs are serving children and families. Important research design features that should be considered include whether the comparison group attended other ECE programs and characteristics of the outcome assessment; specifically, the content, method of assessment, and reliability of the instruments. Facing scarce resources and difficult funding decisions, policymakers would benefit greatly from the ability to compare the costs of different programs relative to their expected benefits across a broad set of school readiness indicators; thus, future work should consider whether and how estimates from diverse evaluations with differing methods and measures can be most effectively compared.

Appendix

Head Start Studies, Programs, and Contrasts Included in the Analysis

Start date	Study description	Programs and contrasts included
1965	Lincoln, Nebraska Summer Head Start	1. Matched pairs: Head Start vs. No Head Start, no preschool (stay at home) 2. Unmatched pairs: Head Start vs. no Head Start, no preschool (stay at home)
1965	Duluth Summer Head Start	1. Head Start vs. no Head Start, no preschool (stay at home)
1965	Westinghouse-Ohio National Head Start Evaluation, 1965–1968	1. Head Start vs. no Head Start, no preschool
1965	Camden, New Jersey Summer Head Start	1. Summer Head Start vs. no Summer Head Start
1965	New Jersey Summer Head Start	1. One or two years of Summer Head Start vs. no Summer Head Start
1965	Cambridge, Massachusetts Summer Head Start	1. Summer Head Start vs. no Summer Head Start and Operation Checkup (medical exam)
1965	Cleveland Summer Head Start	1. Summer Head Start vs. no Head Start
1965	Kearney, Nebraska Summer Head Start	1. Summer Head Start vs. no Summer Head Start
1965	San Jose, California Summer Head Start	1. Summer Head Start vs. no Summer Head Start
1966[a]	New Haven Head Start Evaluation, Smaller Follow-up	1. Head Start vs. no Head Start
1966[a]	New Haven Head Start Evaluation, Larger Follow-up	1. Head Start vs. no Head Start
1966	Washington, DC Summer Head Start	1. Summer Head Start vs. no Head Start
1966[a]	Dade County, Florida Head Start Program Study (effects on self-concept, social skills, and language skills)	1. Head Start vs. no Head Start, no preschool
1966	Bicultural Preschool Program (Mexican American children)	1. Head Start vs. no Head Start, no preschool
1967[a]	New York City Head Start	1. Head Start vs. children about to enter Head Start
1967	Head Start, UCLA Evaluation	1. Head Start vs. no Head Start
1968	Rural Minnesota Head Start	1. Head Start vs. no Head Start, but eligible for Head Start 2. Head Start vs. no Head Start, no preschool, and not eligible for Head Start
1968	Louisville Head Start Curriculum Comparison	1. Bereiter–Engelmann Head Start vs. no Head Start, no preschool 2. DARCEE Head Start vs. no Head Start, no preschool 3. Montessori Head Start vs. no Head Start, no preschool 4. Traditional Head Start vs. no Head Start, no preschool
1968	Evaluation of Standard Head Start and Direct Instruction Head Start	1. Head Start with Bereiter–Engelmann curriculum (direct instruction) vs. no Head Start, no preschool 2. Standard Head Start vs. no Head Start, no preschool
1969	Educational Testing Services Longitudinal Head Start Evaluation (Portland, Oregon and Trenton, New Jersey)	1. Head Start vs. no Head Start, no preschool
1971	Planned Variation Head Start Study	1. Head Start with an added formal approach or curriculum vs. no Head Start, no preschool 2. Standard Head Start vs. no Head Start, no preschool

(continued)

Appendix (continued)

Start date	Study description	Programs and contrasts included
1979	Head Start Bilingual Bicultural Curriculum Models Project	1. Bilingual Bicultural Head Start vs. no Head Start, no preschool (stay at home) 2. Standard Head Start vs. no Head Start, no preschool (stay at home)
1980[a]	New Haven Public-School-Sponsored Head Start	1. Head Start vs. no Head Start, no preschool
1985	Guam Head Start Evaluation	1. Head Start vs. no Head Start
1997	Early Childhood Longitudinal Study–Kindergarten Cohort Head Start Study	1. White children: Head Start vs. no Head Start (mix of stay at home, other preschool, and child care) 2. Black children: Head Start vs. no Head Start (mix of stay at home, other preschool, and child care) 3. Hispanic children: Head Start vs. no Head Start (mix of stay at home, other preschool, and child care)
1998	Southeastern Head Start program of high quality	1. Head Start vs. Head Start wait list
2002	National Head Start Impact Study, First Year	1. Three-year-olds, intent to treat ordinary least squares (weighted, controlling for demographics and pretest scores): Head Start vs. no Head Start (includes crossovers) 2. Four-year-olds, intent to treat ordinary least squares (weighted, controlling for demographics and pretest scores): Head Start vs. no Head Start 3. Three-year-olds, treatment on treated (Ludwig & Phillips analysis): Head Start vs. no Head Start 4. Four-year-olds, treatment on treated (Ludwig & Phillips analysis): Head Start vs. no Head Start

Note. Unless otherwise specified, Head Start refers to a full-year academic program. Contrasts within a single program are numbered. Please see the following website for a full list of references for these studies: http://developingchild.harvard.edu/download_file/-/view_inline/1203/.

a. If an actual start date for the program was not provided, we estimated the start date to be 2 years prior to report publication.

Funding

The authors disclosed receipt of the following financial support for the research and/or authorship of this article: We are grateful to the following funders of the National Forum on Early Childhood Policy and Programs: the Birth to Five Policy Alliance, the Buffett Early Childhood Fund, Casey Family Programs, the McCormick Tribune Foundation, the Norlien Foundation, Harvard University, and an Anonymous Donor. We are also grateful to the Institute of Education Sciences, US Department of Education for supporting this research (#R305A110035), to Abt Associates, Inc. and the National Institute for Early Education Research for making their data available to us. Shager's work on this project was supported by the Institute of Education Sciences, U.S. Department of Education grant to the University of Wisconsin-Madison (#R305C050055).

Note

1. Categories were constructed based on quartile scores and defined as follows: high reliability (reference category), greater than or equal to .92 (representing Quartile 4); medium reliability, .91 to .75 (representing Quartiles 2 and 3); low reliability, less than .75, (representing Quartile 1). Our preference was to code any reliability coefficient provided for the specific study population; however, this information was rarely reported. If no coefficient was provided in the report, we attempted to find a reliability estimate from test manuals or another study. Any available type of reliability coefficient was recorded, although most were measures of internal consistency (Cronbach's alpha). Because of this variability in source information and coefficient type, and the fact that we were still left with missing reliability coefficients for 38% of our effect sizes, we offer these results with caution.

References

Besharov, D. J., & Higney, C. A. (2007). Response to Barnett and Currie. *Journal of Policy Analysis and Management, 26*(3), 686–688.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). Comprehensive Meta-Analysis (Version 2) [Computer software]. Englewood, NJ: Biostat.

Camilli, G., Vargas, S., Ryan, S., & Barnett, W. S. (2010). Meta-analysis of the effects of early education interventions on cognitive and social development. *Teachers College Record, 112*(3), 579–620.

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.

Cook, T. (2006). *What works in publicly funded pre-kindergarten education?* Presented at the Children's Achievement: What the Evidence Says about Teachers, Pre-K Programs and Economic Policies Policy Briefing, Washington, D.C.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27*(4), 724–750.

Cook, T. D., & Wong, V. C. (2007). The warrant for universal pre-K: Can several thin reeds make a strong policy boat? *Social Policy Report, 21*(3), 14–15.

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 ed., pp. 3–17). New York, NY: Russell Sage.

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

de la Torre, J., Camilli, G., Vargas, S., & Vernon, R. F. (2007). Illustration of a multilevel model for meta-analysis. *Measurement and Evaluation in Counseling and Development, 40*, 169–180.

Gibbs, C., Ludwig, J., & Miller, D. (2012). *Does Head Start do any lasting good?* (Working Paper 17452). Cambridge, MA: National Bureau of Economic Research.

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.

Gormley, W. T., Jr., Phillips, D., & Gayer, T. (2008). Preschool programs can boost school readiness. *Science, 320,* 1723–1724.

Henry, G. T., Gordon, C. S., & Rickman, D. K. (2006). Early education policy alternatives: Comparing quality and outcomes of Head Start and state prekindergarten. *Educational Evaluation and Policy Analysis, 28*, 77–99.

Hoyt, W. T., & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4*(4), 403–424.

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

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.

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

Ludwig, J., & Phillips, D. (2007). The benefits and costs of Head Start. *Social Policy Report, 21*(3), 3–18.

Magnuson, K., & Shager, H. (2010). Early education: Progress and promise for children from low-income families. *Children and Youth Services Review, 32*(9), 1186-1198.

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, DC: CSR.

Moher, D., Pham, B., Jones, A., Cook, D. J., Jadad, A. R., Moher, M., . . . Klassen, T. P. (1998). Does quality of reports of randomized trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet, 352*, 609–613.

National Forum on Early Childhood Policy and Programs. (2010). *Understanding the Head Start Impact Study*. Retrieved from http://www.developingchild.harvard.edu/

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.

Rosenshine, B. (2010, March). *Researcher-developed tests and standardized tests: A review of ten meta-analyses*. Paper presented at the Society for Research on Educational Effectiveness conference, Washington, DC.

Rothstein, H. R., & Hopewell, S. (2009). Grey literature. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 103–125). New York, NY: Russell Sage.

Schulz, K. F., Chalmers, I., Hayes, R. J., & Altman, D. G. (1995). Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of the American Medical Association, 273*(5), 408–412.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs* (report from the Governing Board of the American Educational Research Association Grants Program). Washington, DC: American Educational Research Association.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, *103*, 1334–1344.

Shadish, W. R., Galindo, R., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A randomized experiment comparing random and cutoff-based assignment. *Psychological Methods, 16*(2), 179–191.

Sweet, M. A., & Appelbaum, M. I. (2004). Is home visiting an effective strategy? A meta-analytic review of home visiting programs for families with young children. *Child Development, 75*(5), 1435–1456.

U.S. Department of Health and Human Services, Administration for Children and Families. (2005, May). *Head Start impact study: First year findings*. Washington, DC: Author.

U.S. Department of Health and Human Services, Administration for Children and Families. (2010, January). *Head Start impact study: Final report.* Washington, DC: Author.

U.S. Department of Health and Human Services, Administration for Children and Families, Office of Head Start. (2010a). *Head Start performance standards & other regulations*. Retrieved from http://www.acf.hhs.gov/programs/ohs/legislation/index.html

U.S. Department of Health and Human Services, Administration for Children and Families, Office of Head Start. (2010b). *Head Start program fact sheet*. Retrieved from http://www.acf.hhs.gov/programs/ohs/about/fy2010.html

Westinghouse Learning Corporation. (1969). The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development. In *A report presented to the Office of Economic Opportunity* (PB 184328). Washington, DC: Distributed by Clearinghouse for Federal Scientific and Technical Information, U.S. Department of Commerce, National Bureau of Standards, Institute for Applied Technology.

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy Analysis and Management, 27,* 122–154.

Zhai, F., Brooks-Gunn, J., & Waldfogel, J. (2010, March). *Head Start and urban children's school readiness: A birth cohort study in 18 cities.* Paper presented at the Society for Research on Educational Effectiveness conference, Washington, DC.

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

Authors

HILARY M. SHAGER is a research analyst at the Wisconsin Department of Children and Families, 201 East Washington Avenue, Second Floor, P.O. Box 8916, Madison, WI 53708-8916, USA; hilary.shager@wisconsin.gov. Her primary research interests include early childhood education and social welfare policy.

HOLLY S. SCHINDLER is an assistant professor in the College of Education at the University of Washington, Miller Hall, Box 353600, Seattle, WA 98195-3600, USA, hschindl@uw.edu. Her research focuses on early childhood and family studies.

KATHERINE A. MAGNUSON is an associate professor of social work at the University of Wisconsin-Madison, and the Associate Director of the Institute for Poverty Research; 1350 University Ave., Madison, WI 53796; kmagnuson@wisc.edu. Her research focuses on early childhood and social welfare policy.

GREG J. DUNCAN is a Distinguished Professor in the School of Education at University of California, Irvine, 2056 Education, Mail code: 5500 Irvine CA 92697; gduncan@uci.edu. He has published extensively on child poverty and the importance of early academic skills, cognitive and emotional self-regulation as well as health in promoting children's eventual success in school and the labor market.

HIROKAZU YOSHIKAWA is the Walter H. Gale Professor of Education at the Harvard Graduate School of Education; yoshikhi@gse.harvard.edu.

CASSANDRA M. D. HART is an assistant professor of education at the University of California, Davis; One Shields Avenue, Davis, CA 95616; cmdhart@ucdavis.edu. Her primary focus is on quantitative evaluations of state and national education policies.

Manuscript received February 17, 2012
First revision received June 30, 2012
Second revision received August 24, 2012
Accepted August 29, 2012