

Abstract Title Page
Not included in page count.

Title: Timing issues with early childhood education programs: how effect sizes vary by starting age, program duration and persistence of effects

Author(s):

Greg J. Duncan, University of California, Irvine Department of Education (gduncan@uci.edu)

James A. Leak, University of California, Irvine Department of Education (jleak@uci.edu)

Weilin Li, University of California, Irvine Department of Education (weilinl@uci.edu)

Katherine Magnuson, University of Wisconsin-Madison (kmagnuson@wisc.edu),

Holly Schindler, Center for the Developing Child, Harvard University
(holly_schindler@harvard.edu)

Hiro Yoshikawa, Harvard Graduate School of Education (yoshikhi@gse.harvard.edu).

Abstract Body

Limit 5 pages single spaced.

Background / Context:

Description of prior research and its intellectual context.

Three interesting timing issues for early childhood education program are intertwined: optimal starting age, optimal program duration and the persistence of impacts. In the case of starting age, “skill begets skill” human capital production models (Cunha & Heckman, 2007) provide a justification for very early intervention, since boosting skills can improve the productivity of later investments. For our country’s universal K-12 schooling “intervention,” this logic suggests that children most ready for kindergarten are best able to profit from the next 13+ years spent in school. But the same logic may apply to preschool investments. If children most ready for an age 4 pre-K program profit the most from them, it may be better to begin boosting children’s skills at age 3 or even earlier to increase the productivity of the age-4 programs.

On the other hand, early investments not followed up with high-quality subsequent investments may produce only ephemeral impacts. In the case of the Perry Preschool intervention (Schweinhart et al., 1993), the large cognitive impacts estimated shortly after the completion of the program had completely disappeared by age 8, although impacts on achievement, attainment and, eventually, crime and earnings persisted.

With starting age and follow-up length held constant, one would expect that longer programs would produce bigger impacts. And indeed, the Gorey (2001) meta-analysis reported that programs with durations in excess of 3 years had larger effects than 1 or 2 year-duration programs. Other studies have largely come to similar conclusions (Barnett & Lamy, 2006).

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

The focus of this paper centers around timing associated with early childhood education programs and interventions using meta-analytic methods. At any given assessment age, a child’s current age equals starting age, plus duration of program, plus years since program ended. Variability in assessment ages across our studies should enable us to identify the separate effects of all three time-related components. Combining these three components within the same analysis allows us to propose the following research questions:

- 1) When is the optimal timing for an intervention during the prenatal to age 5 period?
- 2) Should early education programs begin shortly after birth or is program initiation at age 3 or 4 just as beneficial for children’s learning?
- 3) Do early programs, which are introduced when children are developing on very different schedules, fade out more quickly than programs introduced later in early childhood?
- 4) Do longer-duration programs have less fade-out than shorter programs?

Setting:

The project is a meta-analysis of evaluation studies of early childhood education programs conducted in the United States and its territories between 1960 and 2007.

Population / Participants / Subjects:

Description of the participants in the study: who, how many, key features or characteristics.

The population of interest is children enrolled in early childhood education programs between the ages of 0 and 5 and their control-group counterparts. Since the data come from a meta-analysis, the population for this study is drawn from many different studies with diverse samples.

Intervention / Program / Practice:

Description of the intervention, program or practice, including details of administration and duration.

Again, as a result of the nature of meta-analysis there is not a particular intervention or program being studied. Instead, we analyze the effects of multiple early childhood education programs including Head Start, Perry Preschool, and many other interventions and early childhood education programs for children ages 0-5.

Research Design:

Description of research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).

The research design of this study is meta-analysis. Instead of students or schools, we use prior studies as our unit of analysis. Meta-analysis allows researchers to gather information about prior studies and then estimate effect sizes of various components of the combined research studies. Effect sizes are expressed in standard deviation units and allow for the aggregation of effects of many programs into an overall program effect (Cooper and Hedges, 2009). Average effect sizes are compared across studies for differences in study design components, domain variables, and other study components.

Data Collection and Analysis:

Description of the methods for collecting and analyzing data.

This project involves a multi-step data collection and evaluation process for determining what studies will be included in the meta-analytic database. The first step was to conduct a comprehensive search of the literature from 1960-2007. The meta-analysis project started in 2007, thus the cut off date for inclusion in the database was 2007. The National Forum on Early Childhood Policy and Programs, which is the umbrella organization for our own work, was able to take advantage of a meta-analytic database compiled by Abt Associates, Inc. and the National Institute for Early Education Research (NIEER), which included early childhood intervention studies from 1960-2003 (Camilli et al., 2010; Jacob, Creps & Boulay, 2004; Layzer, Goodson, Bernstein & Price, 2001). This previous meta-analysis yielded 624 previously coded studies¹. Next we conducted keyword searches in ERIC, PsycINFO, EconLit, and Dissertation Abstracts databases, resulting in 9,617 documents, which we refer to as reports (a particular evaluation may consist of a series of reports). Next, we manually searched the websites of policy institutes (e.g., RAND, Mathematica, NIEER) and state and federal departments (e.g., U.S. Department of Health and Human Services), as well as references mentioned in collected studies and other key early childhood education reviews. This search resulted in another 692 possible reports for inclusion in the database. In sum, 10,309 reports for possible inclusion in the early childhood

¹ The original Abt database included ECE programs evaluated between 1960 and 2003 and used similar search techniques; therefore, we did not re-search for evaluations conducted during these years, with the exception of 2003. We conducted searches for evaluations completed between 2003 and 2007. However, our search process did result in the identification of several evaluations published prior to 2002 that were not included in the Abt database.

education portion of our database were identified, in addition to the 624 previously coded by Abt and NIEER.

Next, we developed criteria for the inclusion of studies into our meta-analytic database. In addition to being an early childhood education intervention or program from 1960 to 2007, studies had to have a treatment and control/comparison group, not simply assessing the growth of one group of children over time. Each of the groups in the study must have included at least 10 participants and incurred less than 50% attrition. Studies were excluded if they were testing a pharmacological agent, assessed children with medical disorders or learning disabilities, or tested the effectiveness of medical procedures or health-related products.

Studies must use random assignment appropriately or one of the following quasi-experimental designs: change models, fixed effects models, regression discontinuity, difference in difference, propensity score matching, interrupted time series, instrumental variables and some other types of matching. Studies that used quasi-experimental designs must have had pre and post test information on the outcome or establish baseline equivalence of groups on several demographic characteristics determined by a joint-test. It was the goal of this meta-analysis to use more rigorous inclusion criteria than previous meta-analyses on this topic and for the quality of quasi-experimental studies that are included to be as close to approximating random assignment as possible.

After preliminary screening of abstracts of early childhood education studies, the vast majority, 91%, of the 10,309 reports were excluded due to violating our inclusion criteria. Most of the excluded reports did not meet the research design criteria, while others were eliminated for methodological errors, or did not meet our eligibility criteria. The resulting database, which is 75% completed, currently contains data from approximately 300 reports. We expect a total of about 400 reports, representing approximately 150 ECE studies for children in programs between birth and age 5 when we complete our coding this winter.

Coding Studies. A team of 9 graduate research assistants (4 at Harvard, 2 at Irvine and 3 at Wisconsin) 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 also passed a reliability test comprised of randomly selected codes from a randomly selected study. In order to pass the reliability test, researchers had to calculate 100% of the effect sizes correctly and achieve 80% agreement 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 involving all four principal investigators, and decisions were kept in an annotated codebook so that decisions about ambiguities could be recalled when coding subsequent studies.

Database. Our database consists of three levels of data: study, contrast, and effect size. Studies are defined as independent investigations of collected data. Contrasts are group comparisons within study (i.e. Head Start vs. non-Head Start, Literacy Intervention vs. no Literacy Intervention, etc...). Effect sizes are comparisons of effects between contrasts on dependent measures which include measures of cognition, achievement, behavior, socio-emotional, and

more. Studies can include multiple contrasts and sub-contrasts and multiple dependent measures. We currently have 162 studies, 882 contrasts and sub-contrasts, and 6,970 effect sizes that are non-missing in our database. We will continue to add studies into early 2011.

Effect Size Computation. This study's outcome measures are reported using effect sizes as the unit of measurement. Effect sizes are computed using the Comprehensive Meta-Analysis computer software program (Borenstein, Hedges, Higgins, & Rothstein, 2005). Hedge's G is the effect size calculation utilized by this meta-analysis. 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 estimator when sample sizes are small.

Measures. Outcome measures for this analysis cover child cognition, achievement, behavior, and socio-emotional outcomes. Cognitive outcomes include measures of theory of the mind, attention, vocabulary, IQ, task persistence, syllabic segmentation such as elision and rhyming. Achievement measures include reading, math, letter recognition, numeracy other than conservation of number, and other achievement tests. Behavior outcomes include Behavior outcomes include health risk behavior, mental health, aggressive / antisocial behavior, delinquency, internalizing, externalizing, developmental disorders, self-esteem, anxious or depressive behavior, withdrawal, impulsive or hyperactive behavior, locus of control. Socio-emotional outcomes include labeling of emotions, delay of gratification/frustration tolerance, pos/neg emotional expression, attachment, social skills, social problem solving.

The independent variables of interest for this analysis include three measures of timing: starting age, length of program, and elapsed time. Starting age is the age of the child at the beginning of the intervention/program. Length of the program is the amount of time (in months or years) that the program lasted. Elapsed time is the time elapsed (in months or years) after the program ended when a follow-up test occurred.

Other variables controlled for in this analysis include measures of reliability, quasi-experimental or random assignment study, attrition, whether the study controlled for baseline measures, activity level of the control group (active or passive), and whether the study was published in a peer reviewed journal.

Data Analysis

Following convention, we express our model in two-level (contrasts within studies and effect sizes within contrasts) hierarchical form which relates effect sizes to the child's: i) age of entry into the program; ii) duration of program and iii) time since the completion of the program. Here the first level of the two-level model is:

$$(1) ES_{ijt} = \beta_{0i} + \beta_{1i}StartAge_{ijt} + \beta_{2i}ProgDuration_{ijt} + \beta_{3i}TimeSinceProgram_{ijt} + \beta_{1i}X_{1ijt} + \dots + \beta_{ki}X_{kijt} + e_{ijt}$$

where the effect size j in contrast i at measurement time t , is modeled as a function of the intercept (β_{0i}), which represents the average (covariate adjusted) effect size for all contrasts, StartAge – the age of the child at the beginning of the program; ProgDuration – the duration of the Head Start program in years; TimeSinceProgram – the number of years between the end of the program and the outcome measurement; the x 's represent measures of program

characteristics, child and family characteristics and study quality measures; and e_{ijt} is a within-contrast error term.

The level-2 equation (contrast level) models the intercept as a function of the grand mean effect size (β_0) and a between-contrast random error term (u_i):

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

To facilitate interpretation of coefficients, all three key variables will be expressed in years. We will experiment with several weighting schemes to take into consideration the within-study variance, within-study dependency, between-study variance, and the sampling error. Specifically, non-iterative and iterative method of moments and weighted hierarchical linear models will be conducted to generate the weighting matrix (Hedges, Tipton, & Johnson, 2010; Stevens & Taylor, 2009; Raudenbush & Bryk, 1985). Since it is unlikely that the effects of each of these variables would be linear, we will experiment with a variety of theoretically-appropriate nonlinear forms. For example, persistence will be estimated using a negative exponential and in more flexible ways using dummy variables. Moreover, we will not assume that the same functional form will fit all outcomes, given, for example, evidence from both Perry and Abecedarian of longer-lived program effects on achievement than IQ. We will also test for such theoretically appropriate interactions such as program duration by time since completion of program to assess whether the longest program have the most enduring impacts.

Findings / Results:

Description of the main findings with specific details.

Given the incomplete nature of our data base, our preliminary data are intended only to provide a rough idea of our likely results. The first two columns of Table 1 provide descriptive statistics on our key timing variables and on other measures we intend to include in our model. Our descriptive show that the average starting age of our programs is at 3.8 years. This will fall as additional studies are added from 0-3 age range. The average length of program is approximately one year and the average follow-up time after treatment is approximately two years. The final column of Table 1 shows regression coefficients and standard errors from a very preliminary model of our timing measures. Significant negative effects are estimated for post-treatment time, which suggests that treatment effects tend to be the highest immediately following the end of treatment. In this model, length of program has an unexpected negative sign, although both this coefficient and the one on starting age are not statistically significant. These results are likely to change as we continue to add data and conduct robustness checks.

Conclusions:

Description of conclusions, recommendations, and limitations based on findings.

Given the preliminary nature of our analysis, we cannot offer conclusions at this point.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

- Barnett, W. S., & Lamy, C. (2006). *Estimated impacts of numbers of years of preschool attendance on vocabulary, literacy and math skills at kindergarten entry*. New Brunswick, NJ: National Institute for Early Education Research.
- 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).
- 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 metaanalysis, 2nd edition*, (pp. 3-17). New York: Russell Sage Foundation.
- Cunha, F., and Heckman, J.J. (2007): The Technology of Skill Formation. *American Economic Review, 97*, 31–47.
- 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.
- Hedges, L. V., Tipton, R., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research synthesis methods, 1*, 39-65.
- 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.
- 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.
- Raudenbush, S.W., & Bryk, A.S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics, 10*(2)75-98.
- Schweinhart, L.J., Barnes, H., and Weikart, D. (1993). *Significant Benefits the High/Scope Perry Preschool Study Through Age 27*. High-Scope Educational Research Foundation, Ypsilanti, MI, Monograph #10.
- Steven, J. R., & Taylor, A. M. (2009). Hierarchical dependence in meta-analysis. *Journal of educational and behavioral statistics, 34*, 46-73.

Appendix B. Tables and Figures

Not included in page count.

Table 1. Preliminary Descriptive Statistics for Starting Time, Duration and Time Since Program

	<i>M</i>	<i>SD</i>	Regression Model
Starting age (yrs)	3.82	1.20	-0.014 (0.025)
Program duration (yrs)	1.04	1.02	-0.044 (0.035)
Time since program (yrs)	1.95	5.43	-.013* (.006)
Number of Effect Sizes	---	---	3609
Number of Contrasts	---	---	300
Number of Studies	---	---	138

Based on partial data. Robust standard errors are given in parentheses. Models are weighted by the inverse variance multiplied by the number of effect sizes within a contrast. Dummy variables included for 4 of 5 domains of interest. Controls for other study design characteristics are also included in the regression model.

$p < 0.05$, * $p < 0.01$, *** $p < 0.001$