Published in final edited form as:

Educ Eval Policy Anal. 2009 November 1; 31(4): 480–499. doi:10.3102/0162373709352239.

Quality of Research Design Moderates Effects of Grade Retention on Achievement: A Meta-analytic, Multi-level Analysis

Chiharu S. Allen, Qi Chen, Victor L. Willson, and Jan N. Hughes

Department of Educational Psychology, Texas A&M University

Abstract

The present meta-analysis examined the effect of grade retention on academic outcomes and investigated systemic sources of variability in effect sizes. Using multi-level modeling, we investigated characteristics of 207 effect sizes across 22 studies published between 1990 and 2007 at two levels: the study (between) and individual (within) levels. Design quality was a study-level variable. Individual level variables were median grade retained and median number of years post retention. Quality of design was associated with less negative effects. Studies employing middle to high methodological designs yielded effect sizes not statistically significantly different from zero and 0.34 higher (more positive) than studies with low design quality. Years post retention was negatively associated with retention effects, and this effect was stronger for studies using grade comparisons versus age comparisons. Results challenge the widely held view that retention has a negative impact on achievement. Suggestions for future research are discussed.

Keywords

Grade retention; Academic achievement; Meta-analysis; Design quality

Grade retention, the practice of requiring a student who has been in a given grade level for a full school year to remain at that level for a subsequent school year (Jackson, 1975), has been a source of controversy in American public education for decades (Bali, Anagnostopoulos, & Roberts, 2005; Lorence, 2006; Owings & Magliaro, 1998). Its use as an educational intervention for low achieving students has fluctuated since the early 1900s, reaching a peak in the 1970s before declining throughout the 1980s and then increasing rapidly in the early 1990s (Bali et al., 20005; Owings & Magliaro, 1998; Wu, West, & Hughes, 2008).

The upsurge in grade retention rates from the 1980s to the mid-1990s has been attributed to the rise of the standards-based reform movement in education (Byrnes & Yamamoto, 1986; Owings & Kaplan, 2001; U. S. Department of Education, 1999) that followed the publication of *A Nation at Risk: The Imperative for Educational Reform* (National Commission on Excellence in Education, 1983). The standards-based reform movement emphasizes setting competency standards for students at each grade level and holding both schools and students accountable for meeting them. The reform movement calls for an end to social promotion, the practice of allowing students who have failed to meet standards to advance to the next grade with their peers instead of completing or satisfying the requirement (U.S. Department of Education, 1999). In his 1998 and 1999 State of the Union addresses, President Clinton urged an end to social promotion and stated that scores on

standardized tests should be the basis for promotion. The *No Child Left Behind* federal legislation passed in 2001 requires that assessments, aligned with state standards, be used to measure the achievement of all children at each grade level (U.S. Department of Education, 2006). A sharp upsurge in the percentage of students retained in promotion gate grades has accompanied implementation of state-level policies that base promotion decisions on performance on tests of grade-level competencies (e.g., Florida Department of Education, 2008; Texas Education Agency, 2007).

Grade retention is an expensive intervention. In 2004, U.S. Census data revealed that 9.6 percent of U.S. youth ages 16-19 had been retained in grade one or more times. Based on the average per pupil expenditure of \$8,916 for the 2004-2005 academic year, the state of Texas spent an estimated 1.7 billion dollars for the extra year of schooling for the 190,802 children retained in grades K-12 in the 2000-2001 academic year (Texas Education Agency, 2006).

Quality of Empirical Evidence of Retention's Effects

Given the prevalence of this expensive educational intervention, one might expect that it enjoys strong empirical support. On the contrary, the empirical evidence of effects of grade retention on both academic and social-emotional adjustment is inconsistent but widely characterized in published literature as negative (for most often cited meta-analyses see Holmes, 1989 and Jimerson, 2001a; for narrative reviews see Jimerson, 2001b; Shepard, Smith, & Marion, 1996; Sipple, Killeen, & Monk, 2004). The negative view of grade retention held by educators is reflected in a policy statement by the National Association of School Psychologists (2003) that “urges schools and parents to seek alternatives to retention that more effectively address the specific instructional needs of academic underachievers.”

Recently researchers have challenged the view that clear conclusions about the effect of grade retention are warranted, based on methodological limitations of extant studies (Hong & Raudenbush, 2005; Wu et al., 2008). For example, Lorence (2006) criticizes prior meta-analytic studies by Holmes (1989) and Jimerson (2001a) for using a “score card” approach to counting the frequency of negative, positive, and non-significant effects or calculating weighted effect sizes without regard for the methodological quality of studies included in the meta-analysis.

Adequacy of Controls for Pre-retention Differences—The most pernicious design issue in studies of the effect of grade retention is making causal inferences about grade retention in the absence of a randomized experimental design (Campbell & Stanley, 1963). For obvious reasons, random assignment of students to the “treatments” of retention and promotion is neither feasible nor ethical. A large number of variables at the child, family, school, and district level are associated with the treatment selection (i.e., retention versus promotion) and with the measured outcomes (Reynolds, 1992; Willson & Hughes, 2006). Due to this potential selection bias, a finding that retained children experience more negative outcomes at some future point in their academic careers, compared to all promoted children, may tell us little about the effect of retention.

In the absence of random assignment, educational researchers have relied on two primary methods to control for pre-retention differences. In the first approach, researchers attempt to control for pre-retention differences by *selecting* a comparable group of low-achieving but promoted students. The most typical method in this approach is to select students in the same grade as the retained students who a) scored below a specific score (e.g., below the 25th or 50th percentile) on a measure of achievement or cognitive ability during that year, but b) were promoted to the next grade the subsequent year. When investigating the impact of retention on academic achievement, minimally the pre-retention measure should be a measure of cognitive ability or achievement, preferably the same measure as the measure of

achievement used to assess the outcome of retention (Lorence, 2006). However, some studies using this approach go further and document equivalent mean group performance for retained and promoted students on pre-retention measures. Unfortunately, even when retained and promoted groups do not differ significantly on pre-retention measures, there is no guarantee that the groups are fully equivalent on the measured variable (Heinsman & Shadish, 1996). More importantly, this approach ignores potential differences between promoted and retained children on other important but not assessed variables, known to be related to school performance (e.g., child hyperactivity, parental education level, peer acceptance).

The second approach, the use of *statistical controls* in analysis of covariance or multiple regression, can compensate for potential selection effects, resulting in a better estimate of the effect of an intervention on an outcome (Campbell & Kenny, 1999). These statistical adjustment procedures assume a) that the limited number of covariates that can be included in the model adequately captures the important pre-existing differences between the retained and promoted groups; b) that the relationship (typically linear) between each covariate and the outcome is correctly specified; and c) that the regression lines for the retained and non-retained groups are parallel. Researchers have rarely reported checking these assumptions, nor have they used alternative procedures which relax one or more of these critical assumptions (see e.g., Little, Hyonggin, Johanns, & Giordani, 2000). In addition, when groups are disparate at pretest (i.e., at pre-retention), the application of statistical adjustment procedures often assumes that the statistical model can be extrapolated *beyond* the region in which the data for the two groups actually overlap, often a very risky procedure (Shadish, Cook, & Campbell, 2002; Shadish, Luellen, & Clark, 2006).

In grade retention literature, these two methods, selecting a comparison group and employing statistical controls, are often used in combination to offset limitations of each approach when used singularly. The use of statistical controls can minimize the problem of non-equivalence between retained and promoted groups at pre-retention on variables that may affect achievement. Similarly, the interpretation of a repeated measures interaction may be qualified by the level of initial control in selection of the groups. In the case of a repeated measures interaction, a completely unselected group interaction may support differential growth, but the association with retention may be weak, as any other variable for which the two groups differ can be an alternative explanation for the effect. Pre-matching on measures correlated with the outcome reduces but does not completely eliminate the regression threat (Hopkins, 1969; Marascuilo & Serlin, 1988).

In an effort to address the adequacy of the comparison group and the quality of the statistical control in meta-analysis literature, Lorence (2006) classified each study that assessed retention effects on achievement in the Holmes (1989) and the Jimerson (2001a) meta-analyses. The comparison group was deemed adequate if the mean pre-retention performance of the promoted group on a measure of ability or on the initial indicator of the outcome was equivalent to the mean pre-retention performance of the retained group. Of the 18 studies in the Jimerson (2001a) study that assessed the effect of retention on achievement, 10 met this criterion of employing an adequate comparison group strategy. The quality of the statistical control was rated as high if the researchers statistically controlled for a pre-retention measure of the outcome variable. Of these same 18 studies, only 4 were rated as using high quality statistical controls. Lorence noted that the few studies reporting positive effects were from better designed studies and argued that educational researchers have been too quick to reach the conclusion that retention is harmful, based on available studies. Although Lorence argued that higher quality studies may produce less negative results, he did not formally test whether design quality of studies systematically accounts for variation in retention effects. Furthermore, because he did not

describe the basis for selection of four recent studies that were not included in previous meta-analyses for review, it is not clear whether his conclusions would hold in a more systematic review of research published since 1990.

Statistical Dependence in Effect Size Estimates—Typically studies of the effect of grade retention on achievement yield more than one effect size. For example, a given published study may report effects on more than one measure of achievement or for more than one grade or for varying years post-retention. Multiple effect sizes, when derived from a single study, are nested within the study; i.e. the individual observations (i.e., effect sizes) are not independent of each other because they are drawn from the same study sample. Thus including all of these effects in the meta-analysis violates the assumption of statistical independence, unless the meta-analysis models the dependence among effect sizes. The statistical effects on estimation, when not accounting for the dependence in some way, are unknown.

One recommended approach to this problem in meta-analysis research is to create an independent set of effect sizes for each outcome construct of interest (e.g., achievement, behavioral adjustment, self-concept) (Lipsey & Wilson, 2001). This is typically achieved by averaging multiple effect sizes for a given construct per study or by selecting only one effect size for a given construct, based on a set of rules. An effect size distribution for each construct can then be constructed and analyzed (Lipsey & Wilson, 2001). This approach, however, while addressing the problem of statistically non-independent effect sizes, is limited in several ways. First, a mean value may not represent or even be similar to any effect in the study. Second, a single value ignores the variability within a study and may imply greater precision for effects than actually occurs. Third, by constraining each study to contribute only one effect size, the researcher is limited in investigating characteristics of variables that differ within studies and may covary with the effects. For example, we were interested in investigating whether effect sizes vary systematically based on the amount of time that had elapsed since the retention year. Within a given study, researchers report effect sizes for multiple time points. Thus, the number of years post-retention is a characteristic of individual effect sizes and not a characteristic of the study. The ability to identify factors that account for variation in effect sizes is greatly limited when one is constrained to reduce effect sizes to one effect per construct per study.

Objectives of This Study

The primary purpose of the current meta-analysis is to identify systematic sources of variability in effect sizes in studies investigating the effect of retention on academic achievement. Using multi-level modeling, we investigate characteristics of effect sizes at two levels: the study (between) level and the individual (within) level. We investigate one study level variable (design quality) and two within-study variables (grade retained and the number of years since the retention intervention).

At study level, all effects within a given study share the same design quality. Consistent with Lorence's (2006) reasoning, we expect that this factor will be associated with less negative effect sizes. With respect to design quality, we expect studies that do a better job of removing the impact of pre-retention differences on outcomes will report more positive (or less negative) effects. It is highly likely that retained and promoted children differ at pre-retention on some variables; otherwise the retained children would have been promoted. We reason that failure to control for these differences through careful selection of the comparison group and through the use of rigorous statistical controls is likely to produce spurious negative effects of grade retention.

We investigate two within-study variables: grade retained and the number of years since the retention intervention. These variables differed across effect sizes within studies. It is commonly believed among educators that retention is more beneficial in the early grades (Tomchin & Impara, 1992). However, recent studies find that the effects of retention in the early grades on long-term adjustment either do not differ from those at later grades (Silbergitt, Jimerson, Burns, & Appleton, 2006) or are more negative (Pagani, Tremblay, Vitaro, Boulerice, & McDuff, 2001). Based on inconsistent findings in the literature, we offer no hypotheses regarding the effect of the grade at which students were retained. However, we expect that whether retention effects become more or less positive with increasing years post-retention may depend on the group comparison strategy used. This expectation is based on evidence that when retained students are compared to their same-grade (and younger) classmates, effects are initially positive but decrease over time (Pierson & Connell, 1992; Wu, West, & Hughes, in press). Conversely, when retained students are compared to their same-age classmates (who are in a grade above), effects are often negative in the short-term but then plateau or become more positive with time (Wu et al., in press). Accordingly, we expect the effect of the length of time since retention on effect sizes will vary based on the group comparison strategy.

Methods

Selection of Studies

Studies were located through computerized databases (e.g., PsycINFO, ERIC, Medline) using subject terms such as grade retention, grade repetition, grade failure, nonpromotion, transition classroom, flunked, and other synonyms. Reference sections of recent review articles were also reviewed to identify relevant articles.

Retention was defined as repeating a grade after having spent a full year in that grade. Placement in developmental kindergarten, pre-first, and transition classrooms was also categorized as retention because students complete one grade over the course of two full academic years. For inclusion in the present analysis, studies had to meet the following criteria:

1. Studies were conducted in North America and published in English between 1990 and June of 2007 in a scholarly book or peer-reviewed journal. We omitted studies prior to 1990 because our interest is on contemporary retention practices. Unpublished conference papers, unpublished technical reports, and doctoral dissertations were excluded. This restriction promotes completeness and accuracy of reporting and aims to increase confidence in the results. (For analysis of publication bias, see *Preliminary Results* below).
2. Studies reported outcomes of retention at the student level. Studies that reported effect of retention as an educational policy were excluded in the review unless the study also provided individual level results. Outcomes had to be reported with quantifiable measures of academic achievement (e.g., test of reading or math or other academic achievement domain; class or report card marks; teacher-rated learning or achievement). Excluded were academically-related variables such as school attendance, post-high school educational or vocational performance, special education placement, or parent-rated or teacher-rated educational expectations.
3. Studies used a quasi-experiment or experimental design with at least one comparison group of similarly achieving students or employed covariates to statistically control for pre-retention differences between retained and promoted children. However, even though we searched for studies using an experimental design, we found no study utilizing this design method. Similarly achieving

students were defined as non-retained students who were similar in their cognitive abilities and/or academic achievement as retained students prior to the repeat year. Studies using a non-selected group of students (e.g., everyone except retained children or a sample taken from everyone but retained children) were excluded from the present analysis, unless the study statistically controlled for pre-retention differences (e.g., Moller, Stearns, Blau, & Land, 2006). Studies were also excluded if the authors provided inadequate information to determine that the covariate employed to statistically control for pre-retention differences had been measured prior to retention (e.g., Roderick, 1994; Silberglitt, Appleton, Burns, & Jimerson, 2006).

4. Studies reported achievement outcomes that did not duplicate previously reported outcomes. This might occur, for example, when the same investigators reported outcomes for the same sample in different publications (e.g., Mantzicopoulos, 1997 and Mantzicopoulos & Morrison, 1992; Reynolds, 1992 and Reynolds & Bezruczko, 1993) or in a subsequent revised edition (Alexander, Entwisle, & Dauber, 1994 and Alexander, Entwisle, & Dauber, 2003). In these cases, we included the publication that used the most rigorous statistical control for pre-retention differences (e.g., Mantzicopoulos, 1997). In the case of equivalent methodological rigor, we included the first publication that reported the results (e.g., Alexander et al., 1994; Reynolds, 1992). However, when two studies reported achievement outcomes at different years post retention, we selected both studies (e.g., Reynolds, 1992; McCoy & Reynolds, 1999).

Of 199 studies that were identified and carefully evaluated as described above, a total of 22 studies met study inclusionary criteria.¹ Descriptive information for the 22 studies included in the current analysis is reported in Tables 1 and 2.

Selection of Groups

It was not uncommon for studies to employ more than one comparison or retention group. If the study employed a statistical method that compared mean differences (e.g., analysis of variance, analysis of covariance), it was necessary to select one comparison group and one retention group for each independent outcome in order to ensure each independent sample contributed only one effect size for a particular outcome. When a study employed more than one comparison group (e.g., a randomly selected group of promoted students and a “low achieving” group of promoted students), we selected the comparison group that had the best quality of match (e.g., Pierson & Connell, 1992; Reynolds, 1992; see *Coding Procedure* below). When a study employed more than one retention group (e.g., students repeating a grade and students placed in pre-first or transition classrooms; Dennebaum & Kulberg, 1994), sample-size weighted statistics were calculated to form one single retention group. Furthermore, weighted statistics were also calculated for studies that reported separate outcomes for nonoverlapping groups such as boys and girls (e.g., Pagani et al., 2001) or black and white students (e.g., McCombs-Thomas et al., 1992).

Selection of Outcomes

Special attention was paid to exclude any duplicate outcomes from the same sample, as some studies contained outcomes that had been previously published. When there were multiple articles that reported the same outcomes on the same subjects, only outcomes reported in the first article were included (e.g., Ferguson, 1991 and Ferguson & Mueller Streib, 1996), given they utilized equivalent statistical control for pre-retention differences.

¹A complete list of 199 articles is available from the first author.

For example, if a first article reported on grade point average (GPA) in 3rd grade and the second article reported on GPAs at the 3rd, 4th and 5th grades, the 3rd grade outcome was selected from the first study and the 4th and 5th grade outcomes were selected from the second study.

Some studies reported both total or composite scores and individual subscale scores on standardized achievement tests (e.g., Johnson, Merrel, & Stover, 1990). If any of the subscale scores were included in the calculation of composite scores, we selected only the subscale scores in order to avoid accounting for overlapping effects. The search produced 22 studies and 207 individual achievement outcomes. For each study, Table 1 reports the number of academic outcomes analyzed.

Coding Procedure

Each study was coded with respect to quality of comparison group, statistical control, and design quality.² Each individual effect was coded with respect to grade retained, number of years post retention, comparison strategy (age or grade comparison), and information necessary to compute effect sizes (see *Calculation of Effect Sizes* below).

Study-Level Coding

Quality of comparison group: At the lowest level of match (1) the comparison group consisted of all promoted students or students who were matched on non-academic variables, such as socioeconomic status (SES). Note that these studies were included in the meta-analysis even though they did not have a low-achieving comparison group because they statistically controlled for pre-retention differences on some variable(s). At the second level of match (2), the comparison group was comprised of low achieving but promoted students. The low achieving students might have performed below some standard, such as 25th percentile on a measure of literacy, or been “recommended for retention” by school personnel but promoted. At the third level of match (3), a finding of no statistically significant differences was reported between the promoted and retained groups on a measure of academic ability or achievement taken prior to retention. While nonsignificance does not imply identical means, it does imply for reasonable sample sizes a high degree of overlap in the distributions of the two groups, which becomes important in evaluating post-retention outcomes. Two of the authors independently rated each study for quality of comparison group ($\kappa = .93$). Disagreements in coding between the raters were resolved through discussion and through re-examination.

Quality of statistical control: Four levels of statistical control were coded. At the lowest level (1), no covariates were employed. At level 2, only distal covariates were employed. Distal covariates were defined as non-achievement and non-ability covariates (e.g., parent educational level or family SES, behavioral conduct, prior attendance in preschool). At level 3, the covariate was a proximal measure of achievement or ability (e.g., score on an achievement test or test of cognitive ability). At level 4, the covariate was the pre-retention measure of the outcome (as in repeated measures designs or in general linear model). Inter-coder agreement was good ($\kappa = .91$). Disagreements in coding between the raters were resolved through discussion and through re-examination.

Design quality was rated as low (1), medium (2) or high (3), based on the joint consideration of the ratings for the comparison group and statistical controls (see Table 3). The reason for considering these two strategies for removing selection differences between the retained and promoted group is that neither strategy is a perfect remedy to non-random assignment, as

²The coding guide is available upon request from the first author.

described above. For example, if the study statistically controlled for the pre-retention measure of the achievement outcome as in repeated measures (statistical control level 4) and the study employed a comparison group that was either selected on the basis of low achievement or ability (quality of match = 2) or was equivalent at pre-retention on a measure of achievement or ability (quality of match = 3), design quality was coded as high or 3 (e.g., Jacob & Lefgren, 2004; Hong & Raudenbush, 2005; respectively). However, even if the study controlled for pre-retention measure of the outcome, if its comparison group was not selected based on low achievement (quality of match = 1), the design quality was coded as medium or 2 (Phelps, Dowdell, Rizzo, Ehrlich, & Wilczenski, 1992), due to concern that the regression slope for the relationship between pretest and posttest within groups may violate the assumption of homogeneity.

If the study statistically controlled for pre-retention ability or achievement differences on a measure different from the outcome measure (statistical control level 3) and employed a comparison group whose mean performance on a measure of ability or achievement was not statistically significantly different from that of the retained students (quality of match = 3), design quality was rated as high or 3 (e.g., Reynolds, 1992). The rationale is that limitations of using a measure of achievement different from the outcome is partly compensated for by the use of a comparison group matched on ability. If a lower level comparison group was used (quality of match = 1 or 2), the design quality was rated as medium or 2 (e.g., Mantzicopoulos, 2003; Alexander et al., 1994).

If the study did not employ statistical controls or employed as covariates non-academic, non-achievement variables (statistical control level 1 or 2), but used a comparison group matched on ability (comparison group level 3), the design quality was rated as medium or 2 (e.g., Pierson & Connell, 1992; Mantzicopoulos & Morrison, 1992; respectively). Otherwise, the study design was rated as low or 1 (e.g., Miesels & Liaw, 1993). The rationale is that, in the absence of statistical controls for achievement or ability differences, employing a comparison group of promoted students who are equivalent to retained students pre-retention on a measure of achievement or ability minimizes, but does not eliminate, the probability that group differences post-retention are a result of third variables that influence both achievement and selection into the retention intervention.

Effect Size-Level Coding

Grade retained: For each effect, we coded the grade at which control students were retained, i.e., grade that they repeated. If an outcome concerned students retained in different grades, a median grade was also determined for each effect. In some studies, exact grades of retention were not clearly indicated. In these cases, attempts were made to contact the authors for clarification. If this effort was not successful, we determined the retention grade(s) based on the information provided in the manuscript and most commonly employed public school practices (e.g., kindergarten through 12th grade). For example, if a study indicated that control students were high school students retained prior to 8th grade (Hagborg, Masella, Palladino, & Shepardson, 1991), we estimated the grades retained to include all grades between kindergarten and 8th grade.

Outcome grade and years post retention: Each effect was first classified as to outcome grade or age. Using this information and grade retained, we determined *Years Post Retention* by comparing the grade retained and outcome grade or outcome age. If a study used an age comparison strategy (see *Comparison Strategy* below), it was necessary for us to convert the age to grade in order to determine the number of years post retention. In these cases, we anchored first-time kindergarten entrance as occurring at age 6, unless more specific information was provided. Furthermore, if a study employed a control group of students

retained at different grades or reported on multiple outcome grades, a median number of years post retention was determined for each effect.

Comparison strategy (grade or age comparison): Each effect was coded with respect to the comparison strategy, i.e., grade or age comparison. In same age comparisons, retained and promoted students are compared on the outcome measure during the same academic year. Typically the retained students will be one grade behind their promoted classmates at the time of comparison. In same grade comparisons, the performance of retained and promoted students is compared when they are in the same grade. This is most commonly conducted by comparing retained and promoted students in the same grade but not in the same year. For example, if the retention decision was made in 1st grade, the performance of promoted students the following year, when they were in 2nd grade, would be compared to the performance of retained students two years later, when they were in 2nd grade. Some studies compared retained and promoted students on grade equivalent scores (e.g., McCoy & Reynolds, 1999; Reynolds, 1992). In this case, the outcome is assessed at the same time (and same age); however, retained and promoted students are compared to the normative performance of students in different grades. Thus, in the example above, the year following students' first year in 1st grade, the performance of promoted students would be compared to grade norms for 2nd graders, whereas the performance of retained students would be compared to grade norms for 1st grade. If a study compared retained students with same-age promoted peers using a grade-equivalent score as an outcome, we coded the effect as employing a grade comparison strategy (for issues associated with grade-equivalent scores when comparing across age, see Lorence, 2006).

Inter-rater agreement (Kappa) for two coders was as follows: grade retained (1.00); outcome grade (.83); years post retention (.83); and comparison strategy (.83). Disagreements in coding between the raters were resolved through discussion and through re-examination.

Based on the effect-level coding, we also determined median grade retained, median outcome grade, and median years post retention across multiple effects for each study, for descriptive purposes only (Table 2).

Calculating Effect Sizes

Effect sizes were estimated using the least approximate (lowest inference) statistics (i.e. most accurate statistics) that were available as suggested by Lipsey and Wilson (2001). Effect sizes were calculated in a consistent manner so that positive scores indicate that the retained group benefitted more than the promoted group and negative scores indicate that the promoted group benefitted more than the retained group. In studies where findings for some outcome measures were reported only as non-significant, the effect size was conservatively estimated as zero. More information on calculation of effect sizes is reported in the appendix.

If sample sizes were equal for each effect within a study, the mean effect for that study is the simple average of the effects. If the effects varied in sample size, a weighted effect was computed using the weight function in Lipsey and Wilson (2001). Thus, there is one effect per study. The overall mean effect for all the studies was computed with one effect per study weighted by the average sample size per study.

Data Analysis

We conducted descriptive and inferential data analyses based on recommendations by Lipsey and Wilson (2001) and on the one effect per study multilevel modeling (MLM) as detailed by Raudenbush and Bryk (2002). In addition to the one-effect-per-study MLM

analysis, we examined studies with several effects using a weighted two-level model with within-study pooled effects to estimate within-study error variance. The study-level weight function (Lipsey & Wilson, 2001) was used to weight study means.

Method Overview

Multi-level regression analyses were conducted using Mplus 4.2 (Muthén & Muthén, 2006a) to test the hypothesized unique effects of design quality, average year out, and median grade retained. To account for the dependence among the observations (effect sizes) within clusters (studies), analyses were conducted using the “two level random” feature of Mplus (version 4.2) that accounts for the nested structure of the data (Muthén & Muthén, 2006a). Between-level sampling weight was used to account for sample size differences (and variance of mean effect differences) between studies (Muthén & Muthén, 2006b, 2006c). The weight was calculated using the Hedges's bias correction formula provided by Lipsey and Wilson (2001),

$$\omega_{sm} = \frac{2n_{G1}n_{G2}(n_{G1}+n_{G2})}{2(n_{G1}+n_{G2})^2 + n_{G1}n_{G2}(ES')^2}$$

where n_{G1} is the number of subjects in Group 1, n_{G2} is the number of subjects in Group 2, and ES' is the unbiased effect size using Hedge's correction in the following formula:

$$ES' = \left[1 - \frac{3}{4N - 9} \right] ES$$

where N is the total sample size ($n_{G1} + n_{G2}$). Note that for some studies, n_{G1} and n_{G2} varied across effect sizes, and the average group sizes across effect sizes were used. We found no meaningful differences between using arithmetic and harmonic means for sample size.

A total of 3 predictors were tested, with 1 at the between level (i.e., design quality) and 2 at the within level (i.e., median grade retained and median year post retention). In the following section, each result will be described.

Results

Descriptive Analyses

Study and Effect Characteristics—A total of 22 separate published studies are included in the present meta-analysis. Table 1 provides study details. Grades of retention varied from kindergarten through 8th grade, with a mean grade of 1.73 ($SD=1.60$). Outcome grades varied from kindergarten through 12th grade, with a mean grade of 4.43 ($SD = 2.90$). The mean number of years post retention was 3.80 years ($SD = 1.94$). The year of retention varied greatly across studies, with the earliest retention occurring in 1979 and the most recent retention occurring in 2004. For the majority of the studies (67%) students were retained in the 1980s. Table 2 displays the variability in the retained grades and outcome grades within study as well as across studies.

A total of 15 studies (68%) employed a grade comparison strategy, while 5 studies (23%) utilized age comparison strategy. Only 2 studies (9%) reported results using both grade and age comparisons. At the effect level, the number of retained students varied from 9 to 4,060 and the number of promoted students varied from 14 to 37,198. At the study level, the

median number of retained students ranged from 17 to 3,203 (mean = 612.69, $SD = 1081.07$) and the median number of promoted students ranged from 17 to 13,420 (mean = 2,911.37, $SD = 6,950.53$). The number of academic outcomes varied from 1 to 108, with a mean of 9.41 ($SD = 22.29$).

Design Quality Characteristics—Concerning the quality of comparison group, 7 studies (32%) were classified as low (coded as 1), 6 studies (27%) as medium (coded as 2), and 9 studies (41%) as high (coded as 3). The mean level of comparison group quality was 2.09 ($SD = 0.87$). In regards to the quality of statistical control, effects were calculated using no covariate in 7 studies (32%; coded as 1), distal covariate in 2 studies (9%; coded as 2), proximal covariate in 6 studies (27%; coded as 3), and repeated measures in 7 studies (32%; coded as 4). Among the 7 studies coded as having the highest level of statistical control, 3 studies did not employ repeated measures in their analyses but provided information that allowed us to conduct our own repeated measure analyses. The mean level of statistical control across studies was 2.59 ($SD = 1.26$). Finally, based on the levels of quality of comparison group and statistical control, we classified the level of design quality as low in 4 studies (18%; coded as 1), medium in 12 studies (55%; coded as 2), and high in 6 studies (27%; coded as 3). The mean level of design control across studies was 2.09 ($SD = 0.68$).

Homogeneity of Effect Sizes and Publication Bias—We also sought to determine the homogeneity of effect sizes and the degree to which results may evidence publication bias. Publication bias refers to a bias toward publishing articles that report statistically significant results over articles that report non-statistically significant results.

Mean effect for each study was plotted against total sample size, and the confidence interval for each effect sample size calculated based on the square root of the inverse of ω_{sm} given above as the standard error, with the mean effect across studies as the population parameter estimate. This produces a so-called funnel plot when the confidence interval limits are connected in more-or-less smooth curves (Light & Pillemer, 1984). The resulting plot indicated five negative effects outside the confidence limits, and six positive effects outside confidence limits. Also, the standard Q chi-square statistic for the study means was computed, $Q = 874.52$, $p < .001$, supportive of heterogeneity in the mean effects. A similar statistic for all 207 effects ignoring study grouping yielded a similar probability.

Taken together, the plot and Q statistic indicate study heterogeneity and a lack of publication bias. We continued our analyses using multi-level modeling to examine effects at both study and effect levels.

Inferential Analyses

Study-Level Analyses—Table 4 lists the predictors used in multi-level analysis, the level at which the predictor was used (“between” means at the study level, whereas “within” means at the individual level within the study), the unstandardized coefficient for the predictor and its standard error, and the p and R^2 values.

Design quality: Two orthogonal contrasts were created to represent the three levels of design quality, with the first contrast representing the difference between the lowest level of design quality and the mean of the other two higher levels, and the second contrast representing the difference between the second and the third higher levels of design quality. The tests of the contrasts indicate that there is substantial difference in effect size magnitude between the first level and the mean of the second and third level ($B = 0.34$, $p = .03$), but there is not a statistically significant difference between the second and the third levels ($B = -0.22$, $p = .45$). This indicates that the two higher levels of design quality lead to less

negative effects and the effect size is on average 0.34 higher (more positive) than studies with the lowest level of design quality. On average, the effect size for studies with the lowest design quality is -0.30, whereas the effect size for studies with medium and high design quality is 0.04.

Effect Size-Level Analyses

Median grade retained: Median grade retained was not significantly correlated with effect size ($B = -0.02, p = .10$). Thus the association between the median grade at which participants in a study were retained and effect size is not statistically significantly different from 0.

Median year post retention: A test including all effect sizes revealed that median year post retention was significantly correlated with effect size ($B = -0.097, p = .009$). Due to the hypothesis that grade and age comparisons might further explain the obtained coefficient, we tested whether the coefficient obtained when pooling all effect sizes would fit the grade or age comparison subgroup of effect sizes. This was done by fixing the path coefficient of the separate analysis for grade and age comparison groups to -0.097, the path coefficient of the combined analysis, because Mplus cannot perform the multiple-group multi-level path analysis. The results showed that the model fit the grade comparison well ($\chi^2(1) = 0.181, p = 0.67, CFI = 1.00, RMSEA = 0.00, SRMR_{\text{between}} = 0.00, SRMR_{\text{within}} = 0.033$), but not the age comparison ($\chi^2(1) = 8.892, p = 0.003, CFI = 0.00, RMSEA = 0.363, SRMR_{\text{between}} = 0.00, SRMR_{\text{within}} = 0.167$).

Furthermore, since the effect size differed based on whether it is based on age or grade comparison, we conducted separate analyses in which we freely estimated the path coefficients of the two comparisons. It was found that in both grade and age comparison, the median year post retention is significantly correlated with effect size ($B_{\text{grade}} = -0.11, p_{\text{grade}} = 0.004; B_{\text{age}} = -0.04, p_{\text{age}} = 0.017$), meaning as the median year post retention increases, the effect size decreases (or become more negative). However, the magnitude of the negative effect for number of years post-retention is greater for grade comparisons (one more year out results in 0.11 decline in effect size) than for age comparisons (one more year out results in only 0.04 decline in effect size).³

Discussion

We assert that our study has extended the current understanding of the effects of retention on academic outcomes through a more careful evaluation of design quality of the existing recent literature. We feel our attention to the type and quality of comparison group permits us to make better conclusions than have heretofore been established. In their seminal work, Glass, McGaw, and Smith (1981) promoted examination of design quality as a major component of a systematic meta-analysis. Their approach, based on Campbell and Stanley's (1963) threats to internal validity, have been extended here to consideration in greater detail the mechanisms for comparison that researchers of retention have made.

Our findings were consistent with study hypotheses. Most importantly, we found that differences in the design quality of studies account for a statistically significant amount of

³Because it is not possible to obtain a true R^2 value for a multi-level model in Mplus (or in HLM), we calculated pseudo R^2 values according to recommendations by Snijders and Bosker (1999). The within-unit variance explained is a measure of how well the independent variables in the model explain the dependent variable. The between-unit measure is the amount of variance between level-2 units that is accounted for by the predictor in the model. Negative estimated R^2 values are possible with this approach, which may result from chance fluctuations or due to misspecification of the larger model. In our case, the negative variance for median grade retained is probably due to chance fluctuations, given the non-significant p value. The negative effect for median year post retention for effects based on grade comparisons is more difficult to interpret and suggests that the R^2 values should be interpreted with caution.

the variability in effect sizes. As expected, studies employing poor methodological controls for non-equivalence at pre-retention between retained and promoted children produced more negative effects of retention on achievement. The mean effect for studies with low design quality was -0.30. An effect of this size falls within the “medium” effect size (Cohen, 1988). Conversely, studies employing better methodological designs (rated as 2 or 3 on a scale based on the joint consideration of the quality of both the comparison group and the statistical controls) had a mean effect size of 0.04, which is not practically or statistically significantly different from 0. That is, studies that do a better job of removing the effect of pre-retention differences on achievement yield a less negative picture of the effects of grade retention. Thus, these results support the view that the null hypothesis of no effect of retention on achievement cannot be rejected. Such a conclusion runs counter to conclusions reached in previously published quantitative and narrative reviews of the retention literature and policy statements based on such reviews (Educational Research Service, 1998; NASP, 2003). It is important to note; however, that these results provide little support for proponents of grade retention. Given the expense of grade retention (Alexander et al., 2003; Foster, 1993; Texas Education Agency, 2006) and the emotional toll retention exacts on students (Anderson, Jimerson, & Whipple, 2005), a finding of “no significant difference” for retention on achievement calls into question the educational benefits of grade retention policies.

On average, effect sizes were assessed 3.80 years post-retention. However, the length of follow-up ranged from 1 year to 13 years. Importantly, with increasing years post-retention, the effects become more negative. Furthermore, as expected, the negative effect of the number of years post-retention is stronger for studies employing grade comparisons than for studies employing age comparisons. When compared to same-grade peers, with each additional year out, on average, the effect of retention on achievement declined by 0.11. When compared to promoted, same age peers, the decline was less steep (decline by 0.04). These results are consistent with the view that retained children experience a short-term boost to their achievement, relative to their younger, same-grade classmates; however, this advantage is lost over the ensuing years. The experience of failure (i.e., being retained in grade), followed by success the following year, followed by a downward trajectory over the ensuing years may negatively effect children's academic self-efficacy and academic control beliefs (Gleason, Kwok, & Hughes, 2007), leading to school disengagement and early school leaving (Alexander et al., 2003). Longitudinal studies that test dynamic models of the effects of early grade retention hold promise for clarifying how retention affects long-term academic and social adaptation. Such an understanding is necessary to realizing the benefits and avoiding the costs of this educational intervention. We did not expect, nor did we find, an association between grade retained and retention effects. Thus our findings do not support the commonly held belief among educators and parents (Tomchin & Impara, 1992) that retention in the early grades is more beneficial, at least with respect to achievement outcomes.

Study Limitations and Future Research Needs

A limitation endemic to meta-analytic studies is the reliance on information reported in articles for analysis. We encountered difficulties with respect to the information needed to calculate effect sizes. During the calculation process, we often needed to estimate statistics such as sample size that are essential for effect size calculation, because such statistics were not available from the published studies and could not be obtained by contacting the authors. Furthermore, we resorted to estimating the effect size conservatively as zero, because the exact test statistics or *p* values was often not reported, when the retention effect was not statistically significant.

Several studies lacked detailed information on pre-retention differences between retained and promoted students or inconsistently controlled for pre-retention differences. For example, Phelps et al. (1992) reported 1st grade achievement scores as measures of pre-retention achievement, even though 35% of their students were retained in kindergarten and 1st grade achievement was measured post retention, not pre-retention. Because we could not control for pre-retention differences in retained and promoted kindergarten children, we excluded their kindergarten sample in our analysis. Similarly, Dennebaum and Kulberg (1994) used school ability index scores measured during the repeat year as an indicator of retained and promoted group equivalence. Future studies should employ only the measures obtained prior to retention to control for pre-retention differences between retained and promoted students.

Future research might also investigate other study features that may moderate the effect of retention on achievement. The degree to which studies rule out differential attrition for retained and promoted students as explanations for retention effects and the psychometric quality of measures of achievement are good candidates for future investigation.

Retention practice varies greatly across school districts and states. As a result, there was a great variability in the practice characteristics. For example, while in many studies students were required to repeat a grade after “failing” a particular grade, in some studies students were placed in pre-grade or transition classrooms prior to going into an age-appropriate grade. The reason for retention also varied from study to study; e.g., from not meeting a state-mandated achievement standard to having low grades on report cards. Furthermore, there was limited report on the additional instructional practices provided during the repeat year. Without examination of this variability, it is impossible to disentangle the effect of grade retention from other unobserved variables pertinent to how retention is practiced. It is likely that retention methodology, defined in terms of how retention is practiced, is associated with its effects on achievement. The present analysis was not able to examine the effects of such variability in practice characteristics.

With respect to the influence of policy context on retention, it may be instructive to examine the four studies conducted in the Chicago Public Schools (CPS) that are included in the current analysis. Two of these studies (McCoy & Reynolds, 1999; Reynolds, 1992) were conducted with the Chicago Longitudinal Sample. The children in this sample were in kindergarten in 1986. The remaining two studies (Jacob & Lefgren, 2004; Roderick & Nagaoka, 2005) were conducted after the CPS implemented standards based accountability practices in the 1996-1997 year. Under the new 1996-1997 policy, students who did not obtain a passing score on the state accountability test in 3rd grade were required to attend a six-week, intensive summer school program, after which they could retake the exam. If they failed the summer exam, they were retained in grade. According to Jacob and Lefgren, the district also provided additional resources to meet the needs of retained students. The effect of retention was small but positive for each of the two studies conducted under the new policy context, compared to moderate and negative effects of retention for each of the earlier studies. These results suggest that retention effects are likely to differ under different policy contexts. Future studies need to incorporate the policy context into their research designs.

The current study analyzed information on the quality of a study's control for potential selection bias into the “treatment” of retention or promotion and determined whether methodological quality moderated retention effects. Obviously, this analysis is constrained by the quality of the studies that are available. It is generally recognized that propensity matching provides the highest quality of control for selection bias, when random assignment is not possible. Traditional methods of adjustment (e.g., covariance adjustment) are limited since they can only use a limited number of observed covariates (Shadish, Cook, &

Campbell, 2002). A propensity score is defined as the conditional probability of assignment to a particular treatment given a vector of observed covariates (Rosenbaum & Rubin, 1983). It is a scalar function of observed covariates that summarizes information required to balance the distribution of the covariates. Propensity score matching is a parsimonious way of reducing bias because it generates a single index—the propensity score—that summarizes information across potential confounds. When propensity scores are used in combination with matching, they provide the strongest level of control for selection effects (Rosenbaum, 2002). Only one study (Hong & Raudenbush, 2005) in the current study used such an approach, and this study's design quality was coded as “3” (High). As more studies on retention effects employ propensity matching (e.g., Wu et al., in press), it will be possible to obtain a better understanding of its effects, independent of pre-retention differences between retained and promoted children.

Conclusions

While offering no justification for the benefits of grade retention, these results challenge the commonly espoused view in published literature that grade retention is negative (Holmes, 1989; Jimerson, 2001a; Jimerson et al., 2006; Shepard et al. 1996). Among studies with minimally adequate controls for pre-retention differences between students who are subsequently retained in grade or promoted, the effect on achievement is not statistically significantly different from zero. The finding that even studies that provide strong controls for selection effects fail to find benefits of grade retention underscores the importance of focusing on the strategies employed when children fail to meet grade level competencies to support the students' academic success. Current estimates of the effects of grade retention are based on current instructional practices for students who struggle academically. Unfortunately, the increase in adoption of grade level promotion gates and other high-stakes accountability strategies has not been accompanied by increases in the instructional supports for struggling students, and too often grade retention just means repeating the prior year's experience (Picklo & Christenson, 2005).

Interestingly, data on the educational services that retained students receive either prior to or during the retention intervention are largely lacking. A recent study compared the average number of “extra” instructional supports (e.g., tutoring, pull-out reading programs, counseling, small group instruction) provided in the year prior to retention and in the year following retention to first grade students who were retained in grade and low-achieving students who were promoted to second grade (Peterson & Hughes, 2008). Controlling for children's propensity scores (i.e., probability of being retained in first grade) constructed from 72 demographic, achievement, and psychosocial variables measured prior to the retention decision, children retained in first grade received fewer instructional supports during their pre-retention year (when all students were in first grade) than did their promoted peers. Furthermore, retained children did not receive more services during the repeat year, relative to the prior year or relative to their promoted peers. These results suggest that retention was not accompanied by increased instructional supports and that the retention may have been considered *the* educational intervention.

Future research on the effects of grade retention needs to focus on the conditions under which repeating a year is beneficial to students, as well as the conditions under which social promotion permits students to “catch up” to their academically more proficient age peers. In particular, prospective, longitudinal research that assesses the provision of instructional supports before and after grade retention or social promotion holds considerable promise for identifying effective educational practices for children who fall below grade level expectations for achievement.

Acknowledgments

This research was supported in part by a grant to Jan N. Hughes from the National Institute of Child Health and Human Development (R01 HD39367).

Appendix

If the study reported a summary statistic for the effect of retention, we converted that statistic to a standardized mean difference effect size, using Lipsey and Wilson's (2001) formulae, with two exceptions. First, in accordance with the principle of least approximation (Lipsey & Wilson, 2001), if the study reported repeated measures data on pre-retention scores on the outcome but did not use these scores in calculating the test statistic, we calculated an effect size based on the interaction term for Time X Group on the outcome measure (e.g. Rust & Wallace, 1993). The second exception occurred when a repeated measures study statistic was based on groups different from the groups of interest to us (e.g. Jimerson, Carlson, Rotert, Egeland, & Sroufe, 1997; Phelps et al., 1992). For example, if a study reported a test statistic based on a low achieving promoted group, all promoted children, and retained children, the mean square within for all groups were used as the estimate of the mean square within groups for the groups of interest, the effect for low achieving promoted versus retained. It is assumed that homogeneity of variance in ANOVA provides an estimate of within group variance suitable for comparisons of the groups of interest. In both cases, we used the reported means and *F* values to calculate the mean square within groups first. Then a simple interaction contrast was computed to obtain a sum of square of the interaction for group by time for each post-treatment time point. Finally the *F* value for interaction was calculated for each post-treatment time point using the calculated mean square within groups and the mean square for interaction and transformed into the equivalent effect size.

In some cases, we had to estimate sample size in order to calculate the standardized mean difference effect size. When only total sample sizes were available (e.g., Jacob & Lefgren, 2004), attempts were made to contact authors for clarification. If this effort was not successful, we used the total sample size divided by two as the best estimate of the sample size for each group. When the sample size for one group was missing (e.g., Alexander et al., 1994), we used the largest possible number based on reported sample characteristics as the estimate of group sample size. When there was no exact chi-square statistic but percentages of each group were reported and the two groups were from independent populations (e.g. Phelps et al., 1992), the arcsine transformation method (Lipsey & Wilson, 2001) was used to estimate the effect. However, when within the same study an exact chi-square value was only available for some outcomes but the percentages of two groups were available for all outcomes (e.g. Ferguson, 1991), the arcsine transformation was used for all effects for consistency within the same study.

References

References marked with an asterisk indicate studies included in the meta-analysis.

- Anderson GE, Jimerson SR, Whipple AD. Student ratings of stressful experiences at home and school: Loss of a parent and grade retention as superlative stressors. *Journal of Applied School Psychology* 2005;21:1–20.
- *. Alexander, K.; Entwisle, D.; Dauber, S. On the success of failure: A reassessment of the effects of retention in the primary grades. New York: Cambridge University Press; 1994.
- Alexander, K.A.; Entwisle, D.R.; Dauber, S.L. On the success of failure: A reassessment of the effects of retention in the primary grades. Cambridge, UK: Cambridge University Press; 2003.

- Bali VA, Anagnostopoulos D, Roberts R. Toward a political explanation of grade retention. *Educational Evaluation and Policy Analysis* 2005;27:133–155.
- Bulla T, Gooden JS. Retention and social promotion: Perspectives of North Carolina elementary school principals. *ERS Spectrum* 2003;21:19–31.
- Byrnes D, Yamamoto K. Views of grade repetition. *Journal of Research and Development in Education* 1986;10:14–20.
- Campbell, DT.; Kenny, DA. A primer on regression artifacts. New York: Guilford Press; 1999.
- Campbell, DT.; Stanley, JC. Experimental and quasi-experimental designs for research. Chicago: Rand McNally; 1963.
- Cohen, J. Statistical power analysis for the behavioral sciences. 2nd. Hillsdale, NJ: Erlbaum; 1988.
- *. Dennebaum JM, Kuhlberg JM. Kindergarten retention and transition classrooms: Their relationship to achievement. *Psychology in the Schools* 1994;31:5–12.
- Educational Research Service. Information for school leaders. Prepared for the Association of California School Administrators. Arlington, VA: Author; 1998.
- *. Ferguson P. Longitudinal outcome differences among promoted and transitional at-risk kindergarten students. *Psychology in the Schools* 1991;28:139–146.
- *. Ferguson P, Mueller Streib M. Longitudinal outcome effects of non-at-risk and at-risk transition first-grade samples: A follow-up study and further analysis. *Psychology in the Schools* 1996;33:38–45.
- Florida Department of Education. Non-promotions in Florida's Public Schools, 2006-2007. 2008. Retrieved October 17, 2008, from <http://www.fldoe.org/eias/eiaspubs/pdf/nonpromotions.pdf>
- Foster J. Retaining children in grade. *Childhood Education* 1993;70:38–43.
- Glass, GV.; McGaw, B.; Smith, ML. Meta-analysis in social research. Beverly Hills, CA: Sage; 1981.
- *. Gleason KA, Kwok OM, Hughes JN. The short-term effect of grade retention on peer relations and academic performance of at-risk first graders. *Elementary School Journal* 2007;107:327–340. [PubMed: 20431696]
- *. Hagborg WJ, Masella G, Palladino P, Shepardson J. A follow-up study of high school students with a history of grade retention. *Psychology in the Schools* 1991;28:310–317.
- Heinsman DT, Shadish WR. Assignment methods in experimentation: When do nonrandomized experiments approximate answers from randomized experiments? *Psychological Methods* 1996;2:154–169.
- Holmes, CT. Grade-level retention effects: A meta-analysis of research studies. In: Shepard, LA.; Smith, ML., editors. *Flunking grades: Research and policies on retention*. London: The Falmer Press; 1989. p. 16-33.
- *. Hong G, Raudenbush SW. Effects of kindergarten retention policy on children's cognitive growth reading and mathematics. *Educational Evaluation and Policy Analysis* 2005;27:205–224.
- Hopkins KD. Regression and the matching fallacy in quasi-experimental research. *Journal of Special Education* 1969;3:329–336.
- Jackson GB. The research evidence on the effects of grade retention. *Review of Educational Research* 1975;45:613–635.
- *. Jacobs BA, Lefgren L. Remedial education and student achievement: A regression-discontinuity analysis. *The Review of Economics and Statistics* 2004;86:226–244.
- Jimerson SR. Meta-analysis of grade retention research: Implications for practice in the 21st century. *School Psychology Review* 2001a;30:420–437.
- Jimerson SR. A synthesis of grade retention research: Looking backward and moving forward. *The California School Psychologist* 2001b;6:47–59.
- Jimerson SR, Pletcher SM, Sarah MW, Graydon K, Schnurr BL, Nickerson AB, et al. Beyond grade retention and social promotion: Promoting the social and academic competence of students. *Psychology in the Schools* 2006;42:85–97.
- *. Jimerson S, Carlson E, Rotert M, Egeland B, Sroufe LA. A prospective, longitudinal study of the correlates and consequences of early grade retention. *Journal of School Psychology* 1997;35:3–25.

- *. Johnson ER, Merrell KW, Stover L. The effects of early grade retention on the academic achievement of fourth-grade students. *Psychology in the Schools* 1990;27:333–338.
- Lipsey, MW.; Wilson, DB. *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications; 2001.
- Light, RJ.; Pillemer, DB. *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press; 1984.
- Little RJ, Hyonggin J, Johannis J, Giordani B. A comparison of subset selection and analysis of covariance for the adjustment of confounders. *Psychological Methods* 2000;5:459–476. [PubMed: 11194208]
- Lorence J. Retention and academic achievement research revisited from a United States perspective. *International Education Journal* 2006;7:731–777.
- *. Lorence J, Dworkin AG. Elementary grade retention in Texas and reading achievement among racial groups: 1994–2002. *Review of Policy Research* 2006;23:999–1033.
- Mantzicopoulos P. Do certain groups of children profit from early grade retention. *Psychology in the Schools* 1997;34:115–127.
- *. Mantzicopoulos P. Academic and school adjustment outcomes following placement in a developmental first-grade program. *Journal of Educational Research* 2003;97:90–105.
- *. Mantzicopoulos P, Morrison D. Kindergarten retention: Academic and behavioral outcomes through the end of the second grade. *American Educational Research Journal* 1992;29:182–198.
- Marascuilo, LA.; Serlin, R. *Statistical methods for the social and behavioral sciences*. New York: W. H. Freeman; 1988.
- *. McCombs-Thomas A, Armistead L, Kempton T, Lynch S, Forehand R, Nousianen S, et al. Early retention: Are there long term beneficial effects? *Psychology in the Schools* 1992;29:342–347.
- *. McCoy AR, Reynolds AJ. Grade retention and school performance: An extended investigation. *Journal of School Psychology* 1999;37:273–298.
- *. Meisels SJ, Liaw FR. Failure in grade: Do retained students catch up? *Journal of Educational Research* 1993;87:69–77.
- *. Moller S, Stearns E, Blau JR, Land KC. Smooth and rough roads to academic achievement: Retention and race/class disparities in high school. *Social Science Research* 2006;35:157–180.
- Muthén, LK.; Muthén, BO. *Mplus user's guide*. Los Angeles, CA: Author; 2006a.
- Muthén, LK.; Muthén, BO. *Mplus user's guide: Version 4.2 language addendum*. Los Angeles, CA: Author; 2006b. Retrieved July 7, 2007, from <http://www.statmodel.com/download/usersguide/Version4.2Language2.pdf>
- Muthén, LK.; Muthén, BO. *Scaling of sampling weights for two-level models*. 2006c. Retrieved July 7, 2007 from <http://www.statmodel.com/download/Scaling3.pdf>
- National Association of School Psychologists. *Position statement on student grade retention and social promotion*. Bethesda, MD: Author; 2003.
- Nacional Commission on Excellence in Education. *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Government Printing Office; 1983.
- No Child Left Behind Act of 2001, Pub L No 107–110. 2002
- Owings WA, Kaplan LS. Standards, retention, and social promotion. *NASSP Bulletin* 2001;85:57–66.
- Owings WA, Magliaro S. Grade retention: A history of failure. *Educational Leadership* 1998;56:86–88.
- Pagani L, Tremblay RE, Vitaro F, Boulerice B, McDuff P. Effect of grade retention on academic performance and behavioral development. *Development and Psychopathology* 2001;13:297–315. [PubMed: 11393648]
- Peterson, LS.; Hughes, JN. Differences between retained and promoted children in educational services received. Paper presented at the American Psychological Association annual conference; Boston, MA. 2008.
- *. Phelps L, Dowdell N, Rizzo FG, Ehrlich P, Wilczenski F. Five to ten years after placement: The long-term efficacy of retention and pre-grade transition. *Journal of Psychoeducational Assessment* 1992;10:116–123.
- Picklo DM, Christenson SL. Alternatives to retention and social promotion: The availability of instructional options. *Remedial and Special Education* 2005;26:258–268.

- *. Pierson LH, Connell JP. Effect of grade retention on self-system processes, school engagement and academic performance. *Journal of Educational Psychology* 1992;84:300–307.
- Raudenbush, SW.; Bryk, AS. *Linear models Applications and data analysis methods*. Thousand Oaks, CA: Sage; 2002.
- Rosenbaum PR. Attributing effects to treatment in matched observational studies. *Journal of the American Statistical Association* 2002;97:183–192.
- Rosenbaum PA, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70:41–55.
- *. Reynolds AJ. Grade retention and school adjustment: An exploratory analysis. *Educational Evaluation and Policy Analysis* 1992;14:101–121.
- Reynolds AJ, Bezruczko N. School adjustment of children at risk through fourth grade. *Merrill-Palmer Quarterly* 1993;39:457–480.
- Roderick M. Grade retention and school dropout: Investigating the association. *American Educational Research Journal* 1994;31:729–759.
- *. Roderick M, Nagaoka J. Retention under Chicago's high-stakes testing program: Helpful, harmful, or harmless? *Educational Evaluation and Policy Analysis* 2005;27:309–340.
- *. Rust J, Wallace K. Effects of grade level retention for four years. *Journal of Instructional Psychology* 1993;20:162–166.
- Shadish, WR.; Cook, TD.; Campbell, DT. *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin; 2002.
- Shadish, WR.; Luellen, JK.; Clark, MH. Propensity scores and quasi-experiments: A testimony to the practical side of Lee Sechrest. In: Bootzin, RR.; McKnight, PE., editors. *Strengthening research methodology: Psychological measurement and evaluation*. Washington, DC: American Psychological Association; 2006. p. 143–157.
- Shepard LA, Smith ML, Marion SF. Failed evidence on grade retention. *Psychology in the Schools* 1996;33:251–261. Review of the book, *On the success of failure: A reassessment of the effects of retention in the primary grades*.
- Silbergliitt B, Appleton JJ, Burns MK, Jimerson SR. Examining the effects of grade retention on student reading performance: A longitudinal study. *Journal of School Psychology* 2006;44:255–270.
- Silbergliitt B, Jimerson SR, Burns MK, Appleton JJ. Does the timing of grade retention make a difference? Examining the effects of early versus late retention. *School Psychology Review* 2006;35:134–141.
- Sipple JW, Killeen K, Monk DH. Adoption and adaptation: School district responses to state imposed learning and graduation requirements. *Educational Evaluation and Policy Analysis* 2004;26:143–168.
- Snijders TAB, Bosker RJ. Standard errors and sample sizes for two-level research. *Journal of Educational Statistics* 1993;18:237–259.
- Texas Education Agency. Texas Education Agency academic excellence indicator system. 2006. Retrieved June 8, 2006, from <http://www.tea.state.tx.us/perfreport/aeis/2001/state.html>
- Texas Education Agency. Austin, TX: Author; 2007. Grade-level retention in Texas public schools, 2005-06 (Document No. GE08 601 01). Retrieved October 17, 2008, from http://www.tea.state.tx.us/research/pdfs/retention_2005-06.pdf
- Tomchin EM, Impara JC. Unraveling teachers' beliefs about grade retention. *American Educational Research Journal* 1992;29:199–123.
- United States Department of Education. Taking responsibility for ending social promotion. 1999. Retrieved on January 23, 2007, from <http://www.ed.gov/PDFDocs/socialprom.pdf>
- United States Department of Education. Improving data quality for Title I standards, assessment, and accountability reporting. 2006. Retrieved on September 23, 2007, from <http://www.ed.gov/policy/elsec/guid/standardsassessment/nclbdataguidance.pdf>
- Willson VL, Hughes JN. Retention of Hispanic/Latino students in first grade: Child, parent, teacher, school, and peer predictors. *Journal of School Psychology* 2006;44:31–49. [PubMed: 20419036]

- Wu W, West SG, Hughes JN. Effect of retention in first grade on children's achievement trajectories over four years: A piecewise growth analysis using propensity score matching. *Journal of Educational Psychology*. in press.
- Wu W, West SG, Hughes JN. Short-term effects of grade retention on the growth rate of Woodcock-Johnson III Broad Math and Reading Scores. *Journal of School Psychology* 2008;46:85–105. [PubMed: 19083352]

Characteristics of Retention Studies

Table 1

Author(s)	Grades Retained	Outcome Grades	Number of Years Post Retention	Grade or Age Comparison	Mean # of Retained Students	Mean # of Promoted Students	# of Academic Outcomes	Comp Group Quality ^a	Stat Control Quality ^b	Design Quality ^c	Mean Hedges Effect Size
Alexander et al. (1994)	1 - 3	1 - 7	1 - 7	Grade & Age	58	68	108	2	3	2	-0.10
Dennenbaum & Kulberg (1994)	K	1 - 3	2 - 4	Grade	17	70	9	1	1	1	-0.54
Ferguson (1991)	K	2	3	Grade	46	20	3	3	1	2	0.85
Ferguson & Mueller-Streib (1996)	K	4	5	Grade	33	14	1	3	1	2	-0.12
Gleason et al. (2007)	1	1	1	Age	63	287	2	2	3	2	0.60
Hagborg et al. (1991)	K - 8	9 - 12	2 - 13	Grade	38	38	3	2	1	1	-0.90
Hong & Raudenbush (2005)	K	K - 1	1 - 2	Age	471	7,168	2	3	4	3	-0.60
Jacob & Lefgren (2004)	3, 6	3, 4, 6, 7	1 - 2	Age	3,170	3,170	8	2	4	3	0.03
Jimerson et al. (1997)	K - 3	1-4, 6, 9	2, 10	Grade	19	32	15	3	4 ^d	3	0.25
Johnson et al. (1990)	K, 1	4	4 - 5	Grade	20	17	5	1	1	1	-0.01
Lorence & Dworkin (2006)	3	3 - 8, 10	1 - 6, 8	Grade	863	29,051	7	2	3	2	0.63
Mantzicopoulos (2003)	K	K - 2	1 - 3	Grade	27	28	6	1	3	2	0.63
Mantzicopoulos & Morrison (1992)	K	K - 2	1 - 3	Grade & Age	53	53	10	3	2	2	0.35
McComb Thomas (1992)	K, 1	2 - 5	2 - 6	Grade	28	28	5	3	1	2	0.00
McCoy & Reynolds (1999)	1 - 7	7	1 - 7	Grade	310	830	4	1	3	2	-0.28
Meisels & Liaw (1993)	K - 8	8	1 - 9	Grade	3,203	13,420	2	1	2	1	-0.30
Moller et al. (2006)	K - 8	8 - 12	1 - 13	Grade	1,805	7,240	1	1	4	2	-0.25
Phelps et al. (1992)	K - 4	7 - 9	4 - 10	Grade	46	24	3	1	4 ^d	2	0.32
Pierson & Connell (1992)	1 - 4	3 - 6	3 - 6	Grade	74	69	1	3	1	2	0.32
Reynolds (1992)	1 - 3	3	1 - 3	Grade	231	200	2	3	3	3	-0.95
Roderick & Nagaoka (2005)	3, 6	3, 4, 6, 7	1 - 2	Age	2,844	2,163	4	2	4	3	-0.01
Rust & Wallace (1993)	K	1 - 3	2 - 4	Age	60	60	6	3	4 ^d	3	0.31
Mean ^e (Standard Deviation ^e)	1.73 (1.60)	4.43 (2.90)	3.80 (1.94)		613 (1,081)	2,911 (6,951)	9.41 (22.29)	2.09 (0.87)	2.59 (1.26)	2.09 (0.68)	-0.11 (0.01)

Note.

^a 1 = All promoted or matched on non-academic variable, 2 = compared to low achieving promoted students, 3 = matched on academic ability or achievement.

*b*₁ = no covariate, 2 = distal covariate, 3 = proximal covariate, 4 = repeated measures or GLM.

*c*₁ = low, 2 = medium, 3 = high.

d Effect sizes are calculated using repeated measures based on the data provided in the manuscript.

e Each study mean effect was weighted by the average sample size per study.

	Piereson & Connell 1992	Alexander et al. 1994	McCoy & Reynolds 1999	Jacob & Lefgren 2006	Roderick & Nagaoka 2005	Meisels & Liaw 1993	Jimerson et al. 1997	Phelps et al. 1992	Lorence & Dworkin 2006	Mantzicopoulos & Morrison 1991	Moller et al. 2006
0.5	2.5	2.0	4.0	3.5	3.5	4.0	1.5	2.0	3.0	0.0	4.0
3.5	4.5	4.0	7.0	5.0	5.0	8.0	5.0	8.0	6.5	1.0	10.0
4.0	4.5	4.0	4.0	1.5	1.5	5.0	6.0	7.0	4.5	2.0	7.0

Table 3
Determination of Quality of Control^a Rating

Quality of Statistical Control	Covariate			Repeated Measures or GLM (4)
Quality of Match	None (1)	Distal (2)	Proximal (3)	
All promoted or matched on non-academic variable (1)	1	1	2	2
Low achieving promoted students (2)	1	1	2	3
Matched on academic ability or achievement (3)	2	2	3	3

Note.

^aQuality of Control - 1: Low; 2: Medium; 3: High.

Table 4
Results of Study-Level and Effect Size-Level Predictors on Achievement Outcomes

Predictor	N	Grade/Age Comparison	Orthogonal Contrast Coding	B	s.e.	p	R ²
Within-Level							
Median Grade Retained	N _{study} = 22 N _{ES} = 207	---	---	-0.022	0.01	0.102	-0.007
Median Year Post Retention	N _{study} = 18 N _{ES} = 147	Grade comparison	---	-0.111	0.04	0.004	-0.128
	N _{study} = 7 N _{ES} = 60	Age comparison	---	-0.040	0.02	0.017	0.461
Between-Level							
Design Quality	N _{study} = 22 N _{ES} = 207	---	c1 (-2/3 1/3 1/3) c2 (0 -1/2 1/2)	0.341 -0.218	0.05 0.15	0.034 0.452	0.395

Note: 1) ES = Effect Size; 2) effect sizes at the within level are nested within studies at the between level; 3) R² is the amount of variance explained by the corresponding predictor at the level in which the predictor was entered (see footnote 3 for more information on the calculation of R²).