International Education Journal, 2006, 7(5), 731-777.
ISSN 1443-1475 © 2006 *Shannon Research Press*.
http://iej.com.au

Retention and academic achievement research revisited from a United States perspective

Jon Lorence
Department of Sociology, University of Houston *jlorencel@uh.edu.usa*

Educational researchers in the United States contend that making low-performing students repeat a grade is an ineffective educational practice. This view derives largely from the summary of grade retention research reported by Holmes (1989). A meta-analysis of more recent studies (Jimerson, 2001) also concludes that the practice of grade retention should be abandoned. However, a thorough examination of the published articles within each of these two meta-analyses reveals that many of the individual studies evidence inadequate research designs and faulty conclusions. The overwhelming majority of conclusions from grade retention studies are unwarranted due to the poor quality of research. Overlooked and more recent retention and grade repartition studies suggest that making students repeat a grade may help increase academic achievement. This review contends that research studies do not support the contention that grade retention is always inappropriate. Suggestions for improving future retention studies are offered.

Grade retention, academic achievement, meta-analysis, faulty conclusions, inadequate research designs, grade repeating

BACKGROUND

The overwhelming majority of educational researchers have concluded that requiring low-achieving students to repeat a grade is an inappropriate, if not harmful, educational practice. Mantzicopoulos and Morrison (1992) contend that "Unlike mixed empirical evidence on other educational issues, research on elementary school nonpromotion is unequivocal. It supports the conclusion that retention is not an effective policy" (p.183). Two of the most vocal opponents of grade retention practices argue that "retention worsens rather than improves the level of student achievement in years following the repeat year" (Shepard and Smith, 1990, p.88). Besides dominating educational research journals, this highly critical view towards making academically-challenged students repeat a grade pervades publications which address the practical concerns of teachers and educational administrators (for example, see Darling-Hammond and Falk, 1997; Harrington-Lueker, 1998; Owings and Magliaro, 1998; Potter, 1996; Reynolds, Temple, and McCoy, 1997). The National Association of School Psychologists view the practice of grade retention to be so ineffective that the organization "urges schools and parents to seek alternatives to retention that more effectively address the specific instructional needs of academic underachievers" (NASP, 2003).

Given such strong beliefs, one would assume that the research demonstrating the futility of grade retention or grade repetition would be equally compelling. However, there is some disagreement regarding the persuasiveness of conclusions commonly associated with grade retention studies. In the most comprehensive review of its time examining the impact of making low-performing students repeat a grade, as opposed to being promoted to the next grade, Jackson (1975, p.627) concluded "that the accumulated research evidence is so poor that valid inferences cannot be

drawn concerning the relative benefits of these two options." More recent meta-analyses of grade retention or grade repetition research (for example, Holmes, 1989; Jimerson, 2001), however, assert that the practice of requiring students to repeat a grade is ineffective at best and likely to be detrimental to further student academic achievement. Conversely, Lorence, Dworkin, Toenjes, and Hill (2002) offered a cursory critique of the grade retention literature and questioned the conclusions of various studies opposed to holding students back in grade for an additional year. Alexander, Entwisle, and Dauber (1994) have been even more critical of the grade retention literature; they argue that an overwhelming number of retention studies are flawed, resulting in erroneous interpretations based on "bad science" (p.220). Shepard (2002, p.57) disagreed with these critical assessment of grade retention studies and accused those questioning the negative impact of grade retention for ignoring the "weight of evidence" which purportedly demonstrates the failure of retention practices. Because educational researchers are likely to be more familiar with the negative conclusions of the grade retention summaries, rather than the specific studies which compose the Holmes and Jimerson meta-analyses, it is worthwhile to explore in greater detail the retention literature. Consequently, readers should not assume that the purpose of this review is to argue for retaining academically challenged students. The major intent of this review is to make readers aware that most of the studies cited in the retention literature are insufficiently sound to support the contention that making students repeat a grade is always wrong.

Although Alexander et al. (2003, chap. 2) point out general shortcomings of the grade retention literature, the current paper presents a more thorough examination of the many specific articles that attempt to assess the effect of retention on student learning outcomes. The purpose of the current paper is to examine the foundation upon which is based the prevailing assertion that requiring students to repeat a grade will not contribute to their academic progress. First specified are criteria used to evaluate the quality of grade retention studies. In addition to usual issues related to the quality of research, a discussion of the appropriate strategy to compare the academic progress of retained and nonretained students is presented. The Holmes (1989) meta-analysis is then scrutinised because his article is often cited as the definitive study which demonstrates that requiring students to repeat a grade is an ineffective educational practice. Next examined is Jimerson's (2001) review which similarly concludes that grade retention policies should be abandoned. An examination of overlooked and new studies pertaining to retention then follows.

CRITERIA OF RESEARCH QUALITY

Published Research

Dunkin (1996) identified nine common types of errors which can occur when attempting to synthesise educational studies (see also Wolf, 1986). Although not all of the problems which can arise in meta-analyses are mentioned here, several errors are pervasive throughout the grade retention literature. The most problematic feature of the retention summaries is that the quality of the individual studies varies considerably; oftentimes mediocre studies are weighted equally with good research. A commonly used indicator of research quality is the type of outlet in which a study appears. Alexander et al. (1994) as well as Dunkin (1996) state that published findings which have undergone a review process are more likely to be of higher quality than unpublished convention papers, master's theses, or dissertations.

Although many readers would likely agree that published findings are probably of higher quality than nonreviewed papers, less consensus exists with respect to other criteria of research quality. Dunkin (1996) warns that assessing the quality of a study "is possibly the most difficult type of error to identify, because criteria for evaluating research are usually controversial, and judges sometimes disagree on the difference between good and poor research" (p.89). To illustrate, critics of grade retention practices assume that the research designs of most studies are appropriate and

yield valid results. For example, Heubert and Hauser (1999, p.122) and Shepard (2000) contend that the Holmes (1989) meta-analysis clearly supports the notion that grade retention is harmful to students because the sum of all the "effects" across the many studies examined are negative. Jimerson (2001) reaches a similar conclusion by counting the number of positive and negative effects calculated from the retention studies he reviewed. His "scorecard" approach yields a greater frequency of negative effects between retained and promoted students. Moreover, average weighted effect sizes Jimerson (2001) presents are also negative. Grade retention is therefore presumed to be an inappropriate educational practice. However, readers seldom question the quality of studies composing these meta-analyses. In addition to the type of publication outlet as an indictor of quality, additional criteria offered to gauge the value of grade retention findings are; research design, type of statistical analysis, comparison strategy, scale of measurement, sample representativeness, and sample size.

RESEARCH DESIGN – MATCHING VERSUS STATISTICAL CONTROLS

Probably the most important criterion for judging the quality of a study is the nature of the research design utilised. Researchers would agree that the best way to assess the effect of different educational practices is by use of an experimental design in which low-performing students are randomly assigned to clearly delineated treatment conditions. Random assignment would help equalise initial differences among students prior to retention which, if ignored, might result in misleading conclusions regarding the effects of grade retention. Observed differences at the end of the study period between retained and promoted students might arise from initial dissimilarities in student abilities rather than retention practices. Although Jackson (1975, p.628) recommended the use of experimental models to assess the effectiveness of grade retention, ethical and practical considerations preclude the implementation of true randomized experimental methods. Most parents and teachers are unwilling to allow random assignment of pupils because of the fear that some students would be harmed if retained (or not retained). Further, random assignment would not be consistent with individual school district promotion policies and practices used to assign pupils to specific grades as well as teachers.

Such constraints have forced researchers to adopt research designs which attempt to approximate the contours of a true randomised experiment. All of the studies cited in the aforementioned meta-analyses, resemble the quasi-experimental "nonequivalent control group design" described by Campbell and Stanley (1963, pp.47-50). In all the retention studies reviewed in this paper, at least two groups of students were studied -- those who were required to repeat a grade and those who were promoted to the next grade. Differences in outcome measures were then examined at the end of the experimental period, usually one or more years after the time of retention.

Controlling by Matching

Due to the inability to randomly assign students to the promoted or retention condition, researchers attempt to equalise initial differences between the two sets of students by the common practice of matching. That is, promoted students with certain characteristics (for example, age, gender, race, and socio-economic standing) presumed to be similar to the retained pupils were identified in school records: Their outcome scores were then compared with those of the retained students. While the strategy of matching is often the only procedure available, most researchers acknowledge that conclusions from such studies should be interpreted cautiously. To illustrate, Campbell and Stanley (1963, p.15) argued that "matching is no real help when used to overcome initial group differences." Campbell and Kenny (1999) later warned: "The danger of matching is that, although the scores are more equivalent due to matching, it is unlikely that they are exactly equivalent. Thus, matching achieves more the illusion of equivalence than the reality" (p.54). With respect to retention practices, Jackson (1975, p.619) pointed out that studies which

compared retained students matched with promoted pupils are biased against those held back because initial differences between the two kinds of students will not be adequately controlled by the practice of matching. Although students may be more similar if matched, promoted pupils will still not be comparable on the most important variables related to retention because those students held back will have greater difficulties in unspecified areas; otherwise they would not have been retained. Alexander et al. (2003) also contend that "It is impossible to match retained and promoted students on all relevant factors – the promoted group, for example, may be more mature, have fewer family problems, or be less aggressive than the held-back group" (p.25). If students are to be equated through matching, a measure of the outcome variable, or at least an indicator of student ability, should be the basis for determining comparability of retained and promoted pupils. In this paper the minimum requirement to be considered appropriately matched is that the two groups should be statistically similar on ability level and/or the initial indicators of the outcome measures. If retained and nonretained students are not comparable, then initial differences should be at least adjusted through statistical methods.

Controlling by Statistical Adjustment

Related to the issue of nonrandom research designs is the choice of appropriate statistical procedures to analyse findings. If one were certain that the retained and promoted students were truly equivalent prior to the year in which pupils repeated a grade, the post-treatment means of the two groups could be compared with a *t* test. Statistically significant differences between groups at the end of the study would then suggest the nature of the effect of retention. Insofar as it is highly unlikely that retained and promoted students are equal in performance at the time of retention, outcome comparisons between groups based on mean differences will be inaccurate and misleading. Campbell and Stanley (1963, p.49) therefore recommended that, when possible, an analysis of covariance be used to help statistically adjust for potential initial differences between nonequivalent groups. More recently Campbell and Kenny (1999, pp.68-79) demonstrated how the use of covariates as statistical controls can help overcome potential selection effects which may result in a better estimate of the impact of a treatment on outcome measures. While many variables could be used for statistical adjustments between groups, it is most important to have a pretreatment measure of the outcome variable because statistically controlling for this pretest will help ensure a truer gauge of the treatment effect, especially if the groups differed on the outcome measure prior to placement in a control or experimental conditions (Maxwell and Delaney, 1990, chap. 9). Other variables which may influence the outcome measure should also be included in the analysis as covariates. As will be seen, only a small number of retention studies focusing on academic achievement attempted to statistically control for initial differences between retained and promoted students, even when such baseline information was available.[1] Although statistical adjustments can yield a more accurate assessment of the effect of making students repeat a grade, traditional linear statistical models can only partially control for initial differences between retained and promoted pupils. The shortcomings of statistical controls will be discussed in more detail after a review of the available retention studies. In spite of problems associated with statistical controls, particularly the availability of variables indicating preexisting differences between retained and nonretained students, many studies cited in the meta-analyses did not adjust for preretention differences even when appropriate variables were available.

[1.] Maxwell, O'Callaghan, and Delaney (1993) speculated that educational researchers' extensive use of matching techniques during the 1970s and 1980s stemmed from a number of articles critical of covariance methods to statistically equalise initial differences between treatment and control groups. However, Maxwell and Delaney (1990, chap. 9) pointed out that some of the earlier criticisms were overly stringent and that analysis of covariance procedures can be useful in helping interpret results from nonrandomized research designs.

GROUP COMPARISON STRATEGIES

Retention studies differ with respect to the time at which retained and nonretained students are compared. Several strategies have been used to assess whether promoted or nonpromoted pupils have advanced relative to one another. Holmes (1989) categorised retention studies as using either a "same-age" or a "same-grade" comparison procedure. The former procedure compares the outcome measure(s) between the retained and promoted students during the same academic year, however, the nonpromoted pupils will be one grade behind their promoted classmates at the time of comparison. Students will be the same age but in a different grade. For example, assume academically challenged first graders were required to repeat the grade. At the end of the repeated first-grade year (that is, the first graders had sat through the same grade twice), the means of retainees would be compared with the average outcome measures of their previous classmates who would be in second grade. Assuming most of the retained and promoted children were of the same age in the initial first grade, the retained students would be of the same age of the nonretained students who had been promoted to second grade. The term "same-year" comparison is also sometimes used to specify that outcome measures are being made between groups of students who are not in the same grade (Karweit, 1992).

Same-grade comparisons, as defined by Holmes (1989, p.21), entails examining the mean performance measures of retained and promoted students when they are in the same grade, but not in the same year. Continuing with the previous scenario, assume the retained first graders had now completed second grade and their initial classmates were one year ahead finishing third grade. Contrasting the second grade scores of the initially retained children with the previous year's means of their promoted classmates, when they were in second grade, results in a same-grade comparison. The time order of the mean outcome measures for the retained students will usually be one year behind that of the promoted pupils in a same-grade comparison. Occasionally, however, the term "same-grade" comparison refers to examining the performance of retained students with their current classmates who had never been required to repeat a grade. The classmates of the retained pupils will usually be one year younger even though they are in the same grade.

The comparison strategy selected for evaluating the impact of grade retention may influence the substantive conclusion of the analysis. Holmes (1989, pp.21-22) summarized that same-age comparisons yielded findings indicating that grade retention results in a negative effect on student outcome measures. Conversely, same-grade comparisons initially support grade retention. However, the positive effect of making students repeat a grade is assumed to quickly diminish in a few years as the gap between those held back and the promoted pupils decreases. Alexander et al. (2003, pp.22-23) speculated that the initial higher scores of retained students observed in same-grade comparisons at the end of the retention year (for example, first grade means of promoted pupils with the second set of first grade means of the students who repeated first grade) occurs because retainees have taken the same curriculum twice. Alternately, promoted students may evidence higher scores than their retained counterparts when performing a same-age comparison because the nonretained students have covered more classroom material due to an additional year of school in a more advanced grade. The promoted students have had an additional academic year of new curriculum than their retained prior classmates.

Given that the basis of comparison may bias findings that support or do not favour retention, it is important to decide whether the same-age or same-grade comparative strategy is substantively more appropriate. Wilson (1990) contends same-grade comparisons are more reasonable because same-age comparisons require retained students to obtain scores comparable to the promoted students, who have been exposed to an additional year of new curriculum. Retained students are in effect required to learn the next year's curriculum in order to obtain the achievement scores

similar to the promoted pupils if a same-age comparison is used. Karweit (1999, pp.43-44) likewise argues that parents and teachers are more concerned with how retained children perform relative to their current classmates rather than their previous peers who were promoted to a different grade. If one assumes that the purpose of making students repeat a grade is to allow them to learn the material necessary to proceed successfully in later grades, then the same-grade comparison is the more appropriate basis for assessing the usefulness of grade retention. Even Shepard (2004), a major critic of retention, agrees that "same-grade comparisons fit better the logic of retention, which is intended to be a one-time adjustment in the student's academic pathway" (p.190). Although some researchers may still prefer same-age contrasts, there is considerable justification that same-grade comparisons are probably more appropriate to assess the academic effect of making students repeat a grade.

SCALE OF MEASUREMENT FOR OUTCOME VARIABLES

A methods issue often ignored in the retention literature is the type of scale or metric used to measure student progress. Standard texts in educational and psychological measurement list numerous procedures to quantify scores (for example, raw scores, normed scores, scaled scores, and grade-equivalent scores). Studies of grade retention often use nationally normed tests to compare the progress of retained and nonretained students. However, the substantive implications of using specific measurement scales have been overlooked by most researchers examining the impact of making students repeat a grade. The use of grade-level or grade-equivalent scores to assess academic achievement is particularly problematic. A grade-equivalent norm is basically the average score obtained by students in a specific grade (Thorndike, 1997, p.60). Methodologists have argued that grade-equivalent scores do not allow adequate measurement of change over time between groups. Analyses demonstrate that the measurement scale used to assess academic achievement can lead to divergent conclusions. For example, Seltzer, Frank, and Bryk (1994) compared growth in Iowa Tests of Basic Skills reading scores over the elementary years among students in Chicago Public Schools. Results based on grade-equivalent scores did not measure growth over time as accurately as logit scores derived from a one-parameter item response theory model. Specifically, grade-equivalent scores should not be used to compare students who are in different grades. Coleman and Karweit (1972) contend:

> A grade-equivalent score, therefore, means a different thing at every grade level. It does not compare the student to others of the same age or at the same actual grade level; it compares him to the average or median student at *another* grade level….[a grade-equivalent score] is not appropriate for inferences about…*rates of growth* of children at different grade levels (pp. 94-95).

Consequently, it is inappropriate to use grade-equivalent measures when students are the same-age but in a different grade, that is the same-age comparison. Retained and promoted students will be in different grades with the nonretained pupils having an additional year of coursework than the retained. Promoted students will therefore likely score an additional grade-level higher than the retained students who will not have covered the same curriculum. Using grade-equivalent scores to compare the same-grade outcomes of retained and promoted students is less problematic because the amount of curriculum covered will be equal for both the promoted and nonpromoted students.

SAMPLE REPRESENTATIVENESS

To generalise findings regarding the effects of grade retention requires that studies incorporate students from different social and economic backgrounds. If researchers examine only one racial or ethnic group, it is unclear if the effects of holding students back a year will affect individuals in other racial/ethnic categories in a similar manner. Likewise, studies based on students only of

European ancestry may yield results inappropriate for students of African American or Hispanic origin who comprise an increasing share of the American educational population. Retention studies should also include economically disadvantaged students as well as those from middle class backgrounds. Some critics of grade retention (for example, Reynolds, 1992) argue that certain kinds of students (specifically poor minorities) are less likely to be helped by retention than White students of higher social standing.

SAMPLE SIZE AND STATISTICAL POWER

In addition to the problem of the representativeness of student samples, Tanner and Galis (1997, pp.109-110) point out that sample sizes are too small in most retention studies to yield valid conclusions. Sample sizes should be large enough to find potential differences between retained and promoted students. Many samples in retention studies lack sufficient power to reject the null hypothesis of no difference between retained and promoted students. Levels of statistical significance and estimated "effect sizes" can be greatly influenced by sample size. Effect size is often reported to indicate the magnitude of the impact of differences in educational practices. Effect size is often calculated as the difference between the mean outcome for an experimental and control group divided by the common standard deviation of the two groups, or the standard deviation of the control group (for example Holmes, 1989). Oftentimes researchers have only 30 pupils or less in either the retained or control groups. The availability of such a small number of students to assess treatment effects results in very low levels of statistical power. Cohen (1977, p.41) demonstrates that analyses with small sample sizes are not large enough to correctly reject the null hypothesis of no difference even if effect sizes are of a moderate magnitude. For example, assume the true effect size is 0.50 (that is, a difference of one half a standard deviation between two groups), which is often defined as a "medium" effect. The power of a *t* test between the two groups each with 30 observations would be 0.61; the probability of concluding that the two means were the same, when in fact they actually differed, would be 0.39. A sample size of 100 students in each group is required to reduce the probability of making a Type II error to 0.06 if the true effect size is 0.50 (Cohen 1977, p.37). Smaller effect sizes require even larger sample sizes to identify statistically significant differences.

In sum, there are various indicators of the quality of research pertaining to grade retention. The publication outlet, the quality of controls to adjust for preretention differences between retained and promoted students, the basis of comparison between the two groups, and the metric of the outcome measures of student achievement. These characteristics of each retention study will be noted and discussed.

PROMINENT META-ANALYSES OF GRADE RETENTION

The aforementioned criteria are used to evaluate the quality of individual research studies cited in the two most prominent reviews on the effectiveness of grade retention. Holmes (1989, hereafter Holmes) updated his earlier review of retention research (Holmes and Matthews, 1984) and aggregated findings from 63 separate studies. Jimerson (2001) reviewed 20 additional articles examining the effects of grade retention on student outcomes. Whereas Holmes reported largely on research conducted between 1960 and 1987, the studies cited by Jimerson (2001) were published during the 1990s.

As previously mentioned, Alexander et al. (1994) and Dunkin (1996) maintain that published articles are usually of a higher quality when compared to unpublished papers and graduate theses or dissertations. Given that Jimerson (2001) also adopted this suggestion, the current paper focuses only on published research findings from journals and books. Ignored in the 1989 Holmes

review are 18 master's theses, 22 dissertations, and two convention papers.[2] These 42 excluded studies comprise 67 per cent of the 63 studies cited in the Holmes meta-analysis. Only one third of the retention studies in the Holmes review were journal articles. Difficulty in acquiring the graduate student documents and convention papers also precluded their inclusion in this review.

The number of studies was further limited by examining only those articles which had findings pertaining to academic achievement. Various researchers have attempted to evaluate the effects of retention on student self-esteem and other psychological outcomes, but the majority of them did not incorporate initial indicators of subjective outcomes (for example, Anfinson, 1941). Given that differences in psychological outcomes between retained and promoted students may have existed prior to the time of retention, it is not possible to assess the impact of making students repeat a grade from many of the studies in the Holmes meta-analyses. Hence, only academic outcomes are considered here: Studies focusing exclusively on subjective dependent variables are ignored. Finally, it was not possible to locate several citations published prior to 1937. These selection criteria left 10 published articles or 16 per cent of those cited in the Holmes (1989) review and 18 out of 20 articles from the Jimerson (2001) meta-analysis of retention studies. The specific articles examined from each of the two meta-analyses are listed in the References.

Holmes Meta-Analysis

Holmes (1989) divided his review into several sections. He first summarized the number of positive and negative differences between the mean outcome measures for retained and promoted students across 63 available studies. He then presented "effect sizes" (mean differences between the retained and promoted pupils divided by the standard deviation of the promoted group) for 25 studies in which the retained and promoted students had been presumably matched or were deemed more similar than in the other 38 studies.[3] Only ten of the 63 studies were refereed published articles. Presently no one has rigorously examined the articles in the Holmes meta-analysis. Dunkin (1996, p.95) commented that "Surely, readers of these works [meta-analyses] cannot go to the trouble of the detailed scrutiny required to check the validity of every synthesis before they decide whether or not to rely on them." The present paper performs such a service by informing readers of the strengths and shortcomings of the articles in the Holmes review. While a discussion of the separate articles may seem tedious, given the frequency that the Holmes meta-analysis is cited (for recent examples see Jimerson, 2004; Shepard, 2004), it is important that readers have some familiarity with the nature of the specific studies in order to determine for themselves the reasonableness of the conclusions inferred from this seminal publication.

Table 1 summarises the major features of the ten published articles cited in the Holmes (1989) meta-analysis. For each article the author and year of publication are listed first. Underneath each reference is the number of students who were required to repeat a grade. The second number specifies the number of promoted pupils in the group whose academic performance is contrasted with the retained students. Several studies include low performing children in addition to students

---

[2] Holmes (1989) cited a convention paper which was later published as a journal article (Shepard and Smith, 1987). Findings from the published paper are presented.

[3] Professor C. T. Holmes (personal communication, April 16, 2004) graciously provided me the name of each study referenced with a number in Table 3 of his chapter on grade retention (Holmes, 1989, p.27). Of the 25 studies Holmes listed as having "matched" subjects, only four were published articles which met the criteria for inclusion in this study. The four studies examined in the current paper which Holmes classified as being of a higher research quality were those by Dobbs and Neville (1967), Peterson, DeGracie, and Ayabe (1987), Sandoval and Fitzgerald (1985), and Shepard and Smith (1987). Nine of the 25 "matched" studies were dissertations, eight were masters' theses, one was an unpublished convention paper, and three articles examined only psychological outcomes.

not experiencing academic difficulties and promoted on a regular basis. Occasionally the numbers of students in these additional comparison groups were unspecified and are denoted by a question mark. The other major characteristics for each article are the quality of quality of controls, the presence of a pretest measure of the outcome variable, an initial measure of student ability, the presence of statistical controls for either an outcome measure or level of initial ability, the nature of the comparison strategy, the use of grade-equivalent scores, and the authors' conclusion regarding the overall usefulness of retention.

Table 1. Characteristics of Retention Studies in Holmes (1989) Meta-Analysis

Author(s)	Quality of Controls	Outcome Pretest	Ability Pretest	Statistical Control for Outcome Pretest	Statistical Control for Ability Pretest	Type of Comparison	Grade Equivalent Units	Authors' Conclusion Effect of Retention
Kamii and Weikert (1963) (31-33)	Inadequate	No	No	No	No	Same- Grade	No	Negative
Sandoval and Fitzgerald (1985) (32-30-75)	Inadequate	No	No	No	No	Same-Grade	No	Negative[b]
Abidin et al. (1971) (85-43)	Inadequate	No	Yes[a]	No	No	Same-Grade	Yes	Negative
Leinhardt (1980) (44-31-?)	Inadequate	No	Yes[a]	No	Yes	Same-Grade Same-Age	No	Negative[b]
May and Welch (1984) (62-59-102)	Inadequate	No	Yes[a]	No	No	Same-Grade	No	Negative[b]
Shepard and Smith (1987) (40-40)	Inadequate?	No	Yes?	No	No	Same-Grade	Yes	Negative
Niklason (1987) (40-62)	Adequate?	Yes	Yes[a]	Yes	No	Same-Age	Yes	Negative
Dobbs and Neville (1967) (24-24)	Adequate?	Yes?	Yes?	No	No	Same-Age	Yes	Negative
Chansky (1964) (30-33)	Adequate	Yesa	Yes[a]	Yes	Yes	Same-Age	Yes	Positive
Peterson et al. (1987) (106-104)	Adequate	Yes	No	No	No	Same-Grade Same-Age	No	Positive

[a] Retained and promoted students were significantly different on this variable at time of retention.
[b] Data indicate that retention had a positive effect on academic achievement.

Quality of controls is denoted as a dichotomous variable as either "adequate" or "inadequate." If the article indicated the retained and promoted students were statistically similar on either an outcome measure at the time of retention or an initial indicator of student ability (for example mental ability) the control procedure was classified as "adequate." Alternately, if the retained and matched promoted students were significantly different on either a measure of initial ability or the initial outcome measure, the quality of the matching is considered to be inadequate. The study is also labelled "inadequate" if indicators of initial ability were available showing retained and promoted students were significantly different and these differences were not statistically controlled. In several articles it was not possible to ascertain if the nonpromoted and promoted students were comparable at the time of retention. Such situations are denoted by a question mark. Of the ten articles, only four meet the criteria for a properly controlled study, and in two of these articles the research designs are questionable. In six of the studies it is highly unlikely that retained and promoted students were similar at the time of retention. Studies can have an adequate research design but still be viewed as "weak" in their ability to assess the effect of grade retention. Small sample sizes, unrepresentative samples, inappropriate grade comparisons, and an inappropriate metric for outcome measures can result in findings with only limited value regarding the impact of retention. An overview of the research characteristics of the ten articles cited in the Holmes meta-analysis are given in Table 1.

One of the first features in Table 1 to raise questions about the extent to which findings from the Holmes meta-analysis can be generalised is the disconcertingly small number of students in the retained and control groups. Only the Peterson, DeGracie, and Ayabe (1987) study examined more than 100 students in the retained and promoted groups. Their analyses, however, were conducted on three separate grade levels (first, second, and third) with samples, respectively, of only 65, 26, and 15 pupils. Small sample sizes of 60 students or less in each group limit the kinds of conclusions which can be made. Note that Chansky (1964) as well as Dobbs and Neville (1967) made inferences based on sample sizes of 30 pupils or less. Admittedly many educational researchers have access to schools with a limited number of students. While investigators can only use data which are available; it is unwise to make broad generalizations about the effectiveness of retention practices based on small numbers of observations. Further, such small samples make it highly unlikely that researchers will be able to find statistically significant differences between retained and socially promoted students. Sampling variation alone could account for the findings. Since the mid-1960s economists and sociologists have required that nonexperimental research examining individual student achievement be based on representative samples with at least several hundred students, if not more, before the findings are taken seriously. Not one of the cited articles meets current standards of sample size requirements for acceptable nonexperimental social science research.

As will become evident, another aspect of these articles limiting their generalisability is that the samples are based on students who attended only a small number of schools in but a few selected school districts. While students from different geographic regions are represented across the different articles, pupils in each study are from only one state. None of the studies provide a student sample representative of the state school population. Were one to require that an adequate research design include a sample size with sufficient statistical power and students who were representative of all pupils enrolled in schools, none of the ten published articles from the Holmes meta-analysis would meet the criteria of an acceptable study. Insofar as some readers may find such criteria too stringent, other characteristics of these studies are examined.

Although Table 1 shows that the authors in eight of the ten studies conclude that making students repeat a grade does not have a positive effect on academic achievement, the overall quality of the research is highly suspect due to inadequate research designs. Table 1 indicates that in only two studies were the retained and promoted students likely comparable in ability or educational progress at the time of retention (Dobbs and Neville, 1967; Peterson et al., 1987). A major problem in interpreting findings from the remaining eight articles is that the nonpromoted and promoted children were of different ability levels at the time of retention, or it was not possible to determine their initial academic progress. Authors were often unable to control for possible differences between retained and promoted students that existed prior to retention. Two studies (Kamii and Weikart, 1963; Sandoval and Fitzgerald, 1985) provided only student outcome measures; no prior information was available to indicate if the retained and promoted students differed in academic ability when students were required to repeat a grade. Kamii and Weikart (1963) examined the educational performance of Michigan seventh graders. Data were obtained retrospectively from the school records of 31 students who had been retained once between first and fifth grade and 31 pupils in the same school district who had never been required to repeat a grade. Measures of student achievement were available only in seventh grade. Analyses using *t* tests revealed that in seventh grade the retained students obtained significantly lower mean reading and arithmetic scores on the Iowa Test of Basic Skills, significantly lower course grades, and significantly lower IQ scores when compared with the promoted students. Although the standard interpretation of such findings is that grade retention does not help students, such a claim is questionable because the promoted students may have demonstrated higher levels of achievement at the time the other students were retained. The lower mean seventh-grade IQ scores

of the retained students suggest that they may have had lower levels of mental ability than the nonretained pupils before retention.

Holmes classified the Sandoval and Fitzgerald (1985) article as being one of the "better matched" studies, but their research design is clearly inadequate. The authors stated that they had matched 75 grade repeaters with students of "comparable ability and motivation" since both sets of students were in the same high school classrooms (Sandoval and Fitzgerald, 1985, p.166). To infer that these California high school students were cognitively similar is highly questionable as students in the same classroom can vary considerably in ability. Retained students in this study had been held back beginning in kindergarten through any other grade prior to high school. To correctly assess the impact of grade retention requires that retained and promoted pupils have similar levels of ability at the time of retention, not years later.

Four other articles in Table 1 designated as having inadequate research designs also provided no initial measures of the outcome indicators between the nonpromoted and promoted students (Abidin, Golladay, and Howerton, 1971; Leinhardt, 1980; May and Welch, 1984; Shepard and Smith, 1987). However, these studies attempted to match student cognitive levels by using tests of students' readiness to learn. Abidin et al. (1971) attempted to control for the ability of retained and promoted students in a south-eastern American school district by matching on the Metropolitan Readiness Test administered to pupils in first grade. Leinhardt (1980) used results from the First Grade Screening Test to identify students for a control group to be compared with retained students. Developmental scores from the Gesell Screening Test were the basis of May and Welch's (1984) identification of retained and promoted kindergartners with similar levels of ability at the time of retention. Shepard and Smith (1987) attempted to match retained and promoted Colorado kindergartners with similar initial levels of developmental ability using the Santa Clara Readiness Inventory. In each of these four studies, however, there is evidence that the nonpromoted and promoted students in the control groups were not of equal ability. Holmes classified the Shepard and Smith article as being one of the better matched studies, but the authors formed matched pairs of children from readiness scores obtained at the ***beginning*** of the kindergarten year in September, not the ***end*** of the year when retention decisions were made. The promoted students could have progressed more rapidly than the retained children during kindergarten, resulting in groups of unequal levels of school readiness at the end of the school year. Shepard and Smith (1987, p.350) further acknowledged that 11 of the 40 matched pairs of students exhibited dissimilar scores on the Santa Clara Readiness test.[4] May and Welch mentioned that the 62 retained kindergartners in a suburban New York state district were more developmentally immature than the control group of 59 children, also recommended for retention but promoted. Neither Shepard and Smith nor May and Welch used the readiness scores as covariates to control for the differences in ability levels between the retained and nonretained children. For these reasons, the research designs of these two studies are classified as inadequate. A more detailed discussion of these four articles follows.

Findings from the Abidin et al. (1971) study are extremely problematic to interpret because of the lack of similarity between promoted and nonpromoted pupils. Although Holmes did not classify the Abidin et al. article as one of the better controlled studies, this paper illustrates the difficulty in reaching a conclusion about the impact of retention when retained and nonretained students with different characteristics can not be properly matched. The authors attempted to equalise the initial

[4] Shepard and Smith (1987) stated that unreported analyses based on the 14 pairs of students who were exactly matched on the Santa Clara Readiness Inventory resulted in effect sizes similar to those presented in the paper. However, it is important to note that the retained and promoted students may not have been similar in ability because they were matched prior to the time of retention.

ability of retained and promoted children in a single school district by matching on scores from the Metropolitan Readiness Test (MRT) taken by first grade students. The academic ability and performance of 85 pupils who had been retained either in first or second grade were compared with 43 children scoring in the bottom quartile of the Metropolitan Readiness Test who had never been retained. Findings from *t* test comparisons between the low-performing promoted and retained students showed the former obtained higher scores on the fourth- and sixth-grade reading and mathematics sections of the SRA. Given that the retained students evidenced higher levels of readiness for school than the nonretained students, the authors concluded that holding students back a year was a "noxious" educational policy. Because many critics of retention (for example, Jimerson, 1999; Meisels and Liaw, 1993; Niklason, 1984; Reynolds, 1992; Southard and May, 1996) cite Abidin et al. as demonstrating the failure of making students repeat a grade, it is worthwhile to thoroughly examine this study.

The data do not support the strong conclusion Abidin et al. (1971) assert. First, retained students and promoted students differed on other characteristics which were uncontrolled. Boys, African Americans, pupils of lower economic standing, and students with working mothers were more likely to be retained, but these variables were not controlled. Other information further indicated the promoted and nonpromoted students were not comparable. Although the authors reported that teachers estimated the retained and nonretained students should have similar academic success, the great dissimilarity in Lorge-Thorndike IQ scores between the two groups suggested that the retained and promoted students were not adequately matched on intellectual ability in first grade. The authors mentioned that the promoted students had significantly higher IQ scores on both the fourth- and sixth-grade tests. Unfortunately no IQ scores were available at the end of first grade, the time retention decisions were initially made for the majority of students, to determine if the retained and nonretained children were of comparable intellectual ability. If one assumes that intelligence levels remain fairly stable over time, the IQ scores of the retained children were likely lower in first grade than the scores of the promoted pupils. The greater intellectual ability of the nonretained children is probably the cause of their higher academic scores in the fourth and sixth grade rather than not being required to repeat a year in grade as Abidin et al. (1971, p.415) claim.

Another factor which limits the interpretability of the Abidin et al. (1971) findings is the reason for grade retention. Besides academic ability, first grade students were retained because of behavioural problems and infrequent school attendance. The observed larger first-grade readiness scores of the retained students may have occurred due to the fact that 15 per cent of the retained pupils were held back because they had not accumulated enough school days to be promoted. It is not unreasonable to speculate that some students who were held back because of excessive absences (perhaps due to illness) could have had higher initial readiness scores than the lower-scoring promoted pupils who did not miss as much school. Deleting from the analyses those students retained for reasons other than academic ability would have helped clarify the relationship between retention and later school success.

The fact that 40 per cent of the retained students (34 of the 85 retainees) were held back in second grade further compromises the matching process because the retained second graders likely evidenced higher first-grade readiness scores than the first graders who were required to repeat that grade. Thirty two per cent of the students repeating the first or second grade were held back because of academic failure, but readers are not informed if the grades of the retained first graders were lower than those of the promoted first graders. Instead, the authors stated that the grades of the retained and promoted groups were the same in first grade.[5] Similarity in subject matter grades

[5] The degree to which the grades of retained and promoted students in first grade differ is confusing. Abidin et al. (1971, p.414) also state: "However, the retained group's grades in both reading and mathematics were significantly lower the year they were retained in the first grade. During the second and third grade there were no significant

in first grade occurred probably because the grades of those students retained later in second grade were comparable to those of the promoted first graders. Combining students retained in first or second grade makes it difficult to gauge the impact of grade retention. A more accurate assessment of retention could have been made if the retained group consisted only of first graders who were held back for academic but not emotional or behavioural reasons.

The four previously mentioned articles tried to match students on the basis of school readiness tests, but research has questioned the usefulness of such tests. Although it seems reasonable to assume a positive correlation exists between school readiness tests and later student academic ability, it is unclear whether readiness tests adequately predict how well students will perform in school. Readiness tests are often so basic (for example, does the student know to read from right to left; can the child recognize certain sounds; does the pupil know major letters, etc.) that they may not be strongly related to academic ability or general intelligence. Several measurement psychologists have argued that the Metropolitan Readiness Tests (MRT) used by Abidin et al. (1971) is not a valid measure of reading ability. Stoner (1995) cited research indicating that the skills assessed by the MRT have relatively little relationship with the ability of beginning learners to read. Mabry (1995) is even more critical of the MRT and stated: "The test is outdated, passé in terms of learning theory, technically inadequate, confusing to targeted audiences, and likely detrimental to children and schools" (p.612). These sentiments may explain why recent studies seldom report use of the Metropolitan Readiness Tests to gauge student ability. The adequacy of the Gesell Test of school readiness used by May and Welch (1984) has also been challenged. Although extensively used in earlier decades, the Gesell tests have significant shortcomings; for example,g., uncertain reliability and validity, inadequate norming samples, and an inability to predict future performance (Kaplan and Saccuzzo, 1997, pp.307-309). Information describing the measurement properties of the Santa Clara Readiness Inventory used by Shepard and Smith (1987) to match retained and nonretained students could not be located.

Unlike the other three studies (Abidin et al., 1971; May and Welch, 1984; Shepard and Smith, 1987) which attempted to control only by matching students on level of school readiness, Leinhardt (1980) used the First Grade Screening Test readiness score in an analysis of covariance to statistically adjust for initial differences in ability levels between retained and promoted children at the end of their kindergarten year. There were 44 pupils whose scores resulted in their being placed into a transition first-grade room (another form of retention) rather than being promoted directly to first grade. The control group consisted of 32 pupils from the previous year's kindergarten class whose screening scores were below the level recommended for promotion to first grade: Nonetheless, these children had been promoted to the first grade. Students in the transition room received individualized reading instruction, as did nine of the 34 at-risk pupils who had been promoted to first grade instead of being placed in the transition room. However, the majority of promoted children experienced a different type of reading program; 23 of the 32 promoted students received basal instruction while all 44 pupils placed in the transition room received New Reading System (NRS) instruction. The screening score and type of reading program were used as covariates to assess the effect of retention on reading performance (measured by the Stanford Achievement Test). Regression results indicated that the retained pupils performed 19 points lower than the promoted kindergartners at the end of 1977 when the promoted students had completed first grade and the nonpromoted pupils were at the end of the transition year. However, by the end of their first grade year, the reading scores of the transitioned students were comparable with those of their first grade classmates who had been promoted to

---

differences in grades between the promoted and retained group nor for the retained group between grades repeated and the original grades."

first grade after kindergarten. The first-grade SAT reading score for the retained students was 77.9 while the end of first-grade reading score of their younger classmates was 60.5, a difference in reading scores of 17.4. This gap was statistically insignificant, probably due to the small sample, but the effect size of 0.97 was large.[6] Had Leinhardt used the initial prescreening test score as a covariate, the difference in first grade reading results between retainees and their year younger classmates might have been even larger. Nonetheless, Holmes (1989, p.26) as well as Heubert and Hauser (1999, p.120), contend that Leinhardt's study demonstrates the ineffectiveness of grade retention.

However, opponents of grade retention ignore Leinhardt's (1980, pp.59-60) discussion which reveals that the major discrepancy between the transition-room pupil reading scores and the first-grade reading scores of the promoted pupils most likely resulted from differences in the quality of reading instruction between the two groups. Pupils in the 1976-1977 transition room were given two and one-half hours ***less*** reading instruction per week than the promoted kindergartners who were in first grade that school year. In addition, the transitioned students received ***less*** than one-half of the test-relevant information given the promoted pupils. The promoted students were also tested more frequently than the kindergartners put in the transition room. The promoted students "were taught the basics of reading directly, more often, and for longer periods of time" (Leinhardt, 1980, p.60). Instead of being given direct instruction in reading, pupils placed in the transition first grade were taught learning skills. The extra help retained students received was inferior to that provided the promoted pupils. Insofar as placing students in a transition room was associated with poorer quality reading instruction, it is not possible to reach any valid conclusions about the effects of a transition first-grade year from this study. The Leinhardt study is inconclusive because the treatment condition (retention) is completely confounded with the nature of reading instruction.

The quality of the research design from the Niklason (1984) article is labelled adequate, although the degree to which initial differences between retained and promoted children were controlled is uncertain (as indicated by a question mark in Table 1). Unlike most of the studies cited by Holmes, Niklason included both an initial indicator of the outcome measure as well as student intellectual functioning in her analysis of urban and suburban Utah elementary school students. Teachers recommended that 144 pupils repeat their current grade (kindergarten through sixth grade). Among the 102 students with available data during the period of study, only 40 students were retained in spite of the teachers' recommendations. Both groups of students were tested at the end of the school year in 1980 and 1981. The 1980 score for either reading or arithmetic from the Wide Range Achievement Test was used as the covariate to predict the 1981 score. Niklason reported promoted pupils showed significantly greater growth in reading achievement than the retained students between 1980 and 1981. The author concluded: "The results of this study…showed that retaining students did not serve the intended purpose of increasing the student's growth academically or in personal or social adjustment" (Niklason, 1984, p.496). This interpretation is problematic because the author also reported that in 1980 the retained students had significantly lower levels of performance ability, along with greater problems in personal and social adjustment, than the promoted children. Unfortunately, Niklason never defined the term "performance ability" but it does not appear to be the same as verbal intellectual ability or academic achievement. Had Niklason included earlier measures of performance ability, personal,

[6] Holmes (1989, p.18) computed the effect size by dividing the mean difference between the retained and promoted students by the standard deviation of the promoted group. For this specific situation, the effect size was calculated as follows: (77.9-60.5)/17.9=0.97. Unless specified otherwise, the same computational procedure is used to estimate effect sizes throughout the paper.

and social adjustment as additional covariates, along with the pretest measure to control for these student differences, the findings would have been more conclusive.[7]

Similar to Niklason (1984), two other articles included indicators of student ability and an initial outcome measure, but these two studies may have better controlled for initial differences between retained and nonretained students. Dobbs and Neville (1967) appeared to better match students on initial levels of academic achievement and ability. They compared 30 first graders who had to repeat the grade with 30 never-retained second graders. Pupils were matched on race, sex, type of classroom, age, mental ability, reading achievement, and school socioeconomic status. Arithmetic scores of the retained students were significantly behind those of the promoted. Beginning in 1962 with reading and arithmetic scores from the Metropolitan Achievement Test, additional test results were obtained at the end of the 1963 and 1964 school years among the remaining 24 students with complete data. The promoted pupils showed significant greater gains in both reading and arithmetic than the retained students. Given the lack of statistically significant differences between the promoted and nonpromoted with respect to age, IQ, and reading level, Dobbs and Neville concluded "that promotion led to the increased achievement gain of the promoted group" (p.474).

Still, the comparability of retained and never-retained students is unclear because the authors did not state if the reading scores for the retained students were taken at the end of their initial year or repeated year of the first grade. Likewise, the test grade year used to match the promoted students with the retained is vague. One presumes that the 1962 achievement scores used to match the two groups of students were the initial first grade test results for both the retained and promoted pupils. Although the initial group means for mental ability and reading achievement were statistically similar, these comparisons were based on 30 students in each group. However, the analyses assessing the change in grades over time were based on only 24 students, as some pupils initially in the study had missing data in later years. A difference of six students may seem minor, but 20 per cent of the initial sample was missing. No information was given indicating if the matched samples of 24 students actually analysed over time were similar in IQ or reading ability. The fact that the IQ difference between the initial 30 retained and promoted students approached statistical significance ($t = 1.50$) suggests that the promoted pupils had higher mental abilities than the children required to repeat first grade. In addition, students may not have been well matched on socioeconomic status because matching was based on school-level rather than individual-level characteristics. Not all children within a school necessarily have families with similar income and social characteristics. These features of the study raise questions about the overall quality of the research. The degree of similarity between the retained and promoted students is uncertain.

Of the two studies reporting that holding students back a grade had a positive effect on academic performance, Chansky (1964) utilized a different statistical adjustment procedure to equalise differences in ability between nonpromoted and promoted low performing children. Thirty first-graders, whose teachers considered the students to be a "poor risk" for success in second grade, were held back, while 33 low-performing students who were judged by their teachers and principals to be a "good risk" were promoted to second grade. First-grade teachers in a rural New York county initially selected 63 students to be retained but 30 were placed into second grade. The promoted pupils scored significantly higher on the California Achievement Test and also

[7] In a later paper Niklason (1987) attempted to control for variables other than the pretest value of the outcome measure. Instead of using the control variables as covariates, however, she created dummy variables and performed an analysis of variance which reduced the power of the control variables. This reanalysis of the Utah data is also problematic because the author used a hierarchical procedure to gauge the effect of grade retention which appears not to have controlled for the other variables later entered into the equation (Niklason, 1987, p.342).

evidenced higher levels of mental ability than those retained at the end of the study period. However, the California Achievement Test enabled computation of an Intellectual Status Index which measured academic performance relative to mental age. Although the promoted students showed slightly greater gains in reading by the end of the next year, comparing observed performance ***relative to expected performance***, indicated that "…the retained group was less discrepant from mental age expectancy than the promoted group" (Chansky, 1964, p.230). Requiring the more academically challenged students to repeat first grade at least enabled them to perform at a level consistent with their mental ability. The promoted students, however, had much lower scores in reading comprehension, arithmetic reasoning, and arithmetic fundamentals than expected, given their intellectual ability. The observed positive effect of retention may partially be a function of the measurement procedure, as this is the only study using a discrepancy technique to control for differences in the mental ability of retained and promoted students.

The only other published study, appearing in the Holmes review, reporting a positive effect for retention examined elementary school students in Mesa, Arizona. Holmes listed this article as having one of the "better matched" groups of students. Given that the retained and nonretained students exhibited similar scores on the outcome measures at the time of retention, the study also met the criteria of adequate controls. Peterson et al. (1987) tracked the academic performance of first, second, and third grade students who repeated a grade. Students were matched on sex, ethnicity, age, and scores from the California Achievement Test (CAT). Separate analyses, by grade, over a three year period revealed that the retained students obtained higher reading, language, and math scores than their promoted counterparts. Although the advantage of first grade retention no longer existed in reading by third grade and mathematics by second grade, the benefits to second and third grade retention persisted over the next three years. The authors concluded that making students repeat a grade resulted in positive academic achievement.[8]

Thus far only four studies have been identified from the Holmes articles which can be classified as having adequate controls for differences between retained and promoted children (that is, Chansky, 1964; Dobbs and Neville, 1967; Niklason, 1984; Peterson et al., 1987), although the assessment is questionable in two of the articles. Holmes did not list either the Chansky (1964) or the Niklason (1984) analyses in his table of "better matched" studies. Whereas the findings in Chansky (1964) and Peterson et al. (1987) favoured making students repeat a grade, the other two studies concluded that retention was unsuccessful in improving student performance. It appears as though the four studies with better controls yield conclusions which, at best, cancel out one another. However, additional questions can be raised about the methodological quality of the two negative studies. Differences between the retained and socially promoted students may not be as well controlled in the Dobbs and Neville (1967) and Niklason (1984) articles as occurs in the other two articles. Further, the two articles with negative conclusions were biased against finding a positive effect for retention given the type of comparisons made. Dobbs and Neville (1964), as well as Niklason (1984), compared retainees and nonretainees when they were in different grades (that is, the same-age contrast). These authors ignored that the scores of the retained students were

8 Alexander et al. (2003, p. 24) suggest that the higher test scores of the retained students relative to the nonretained children may partially be due to the use of a measurement scale based on normal curve equivalents (NCEs). Whereas the means of the retained first graders were based on NCEs normed to first grade students, the means of the second graders would have been scaled relative to the distribution of all second graders used to calculate the national norms. In short, the means of the retainees and the nonretained students were not based on the same reference group used to norm their raw scores. This lack of a common reference point may be problematic for Peterson's et al. (1987) same-year analyses but the same-grade mean contrasts are likely based on norms derived from the same set of pupils. For example, the first grade means of both the retained and promoted pupils are probably derived from the same set of first graders throughout the nation used to establish the first grade norms.

one grade behind the promoted children, who had experienced an additional year of new curriculum than those held back. The higher scores of the socially promoted students may derive from their having an additional year of being introduced to more advanced material than the retainees. Another feature of the two analyses favouring the promoted pupils was that the outcome measures were based on grade-equivalent units. Given that they had not been introduced to the next year's curriculum, the retained students were less likely to score as high as the promoted students, who were one grade ahead and had received an additional year's worth of instruction of more advanced educational material.

Although only two of the ten reviewed articles from the Holmes (1989) meta-analysis concluded that making students repeat a grade was worthwhile, findings from several of the other articles can be interpreted to support retention. For example, even though they believed retention in later grades was ineffective, Sandoval and Fitzgerald (1985, p.169) stated that children who attended a first-grade transition room between kindergarten and first grade, rather than being promoted, demonstrated academic performance superior to their "matched" senior high classmates. Proponents of retention argue that making students repeat a grade will help them learn the material needed to catch-up with classmates who did not have academic difficulties. Two articles indicated that students who had been struggling with subject matter learned enough after a year of retention to reach a level of academic achievement comparable with nonretained peers. Leinhardt (1980) reported that the retained students had higher first-grade Stanford reading scores than their classmates who had not been placed in a transition room. May and Welch (1984) found that kindergartners who had been required to repeat a grade because of low-developmental readiness scores later performed as well as other developmentally immature students who had been socially promoted. The mean third grade New York Pupil Evaluation Test scores were similar between the retained and socially promoted, as were their scores on the Stanford Achievement Tests. This study concluded retention was ineffective because the authors assumed the retained students should have had higher scores than their socially promoted peers who had experienced one less year of school. But May and Welch overlooked the fact that the retained students were initially significantly less mature than the socially promoted students (p.384). One could assume that a transition year helped the lower-performing children catch up with their nonretained peers. The absence of a statistically significant difference between promoted and nonpromoted pupils can be interpreted to support an advantage for retention if the performance of the latter students was appreciably below that of the nonretained children at the time of retention.

Critics of grade retention often cite mean "effect size" differences between retained and nonretained students to conclude that making students repeat a grade is ineffective. Holmes (1989, p.27) reports an average effect size of -0.28, based on the 25 studies he classified as "best-matched. Even if only the 16 studies which had IQ or achievement scores which presumably measured initial student ability are considered, the effect size is -0.30. However, the term "effect size" is a misnomer, particularly in the retention literature (and probably in other substantive areas), because ***effect*** denotes a causal change or the impact of a treatment on an outcome measure. It must be remembered that an "effect size" is merely a mathematical term indicating only how many standard deviation units two means differ. The important question is not the average magnitude of the differences, but what are the factors which ***explain*** the observed differences? Effect size can not be interpreted to imply retention causes an outcome, unless alternative explanations can reasonably be ruled out. The average effect size which can be calculated from the higher quality studies in Holmes varies by which studies are included in the analyses. For example, Shepard, Smith, and Marion (1996, p.252) reported the average effect size from the "six most tightly controlled studies showed retained students behind the controls by one-

quarter of a standard deviation."[9] However, if the Peterson et al. (1987) study, which controls for initial achievement level, SES, and other characteristics, is included with the six studies used by Shepard et al. (1996), the average effect size is reduced in half to -0.10.

These calculations indicate that the choice of variables to be controlled also influences the estimated average effect size. Assume that one were to predict an academic outcome by controlling for initial scores on an achievement test along with socioeconomic status. Holmes (1989, p.27) listed only five studies among the 25 studies with students considered matched on both initial achievement score and SES.[10] The average effect size for these studies is a greatly reduced -0.01. Thus there is no meaningful difference between retained and promoted students after taking into consideration initial level of academic achievement and social-economic background. If the Dobbs and Neville (1967) article were deleted from the computations because of the uncertainty of the matching process, the average effect size would be 0.12, indicating that retention results in weak positive academic benefits. In sum, the impact of grade retention on student performance depends on which variables are considered more important as controls and how well retained and promoted students are actually matched.

The indiscriminate use of these reported effect sizes would lead one to argue that retention does no good, but such an interpretation is valid only if the two sets of students were similar at the time of retention. As has been pointed out, it is unclear whether retained and socially promoted students in the Dobbs and Neville (1967) or Shepard and Smith (1987) study were truly similar in ability at the time of retention. There are other variables that should be controlled in the cited analyses. For example, retained and nonretained students may differ with respect to parental support. Several studies (May and Welch, 1984; Niklason, 1984; Shepard and Smith, 1987) noted the low-performing students in the control groups were often socially promoted at the request of their parents. Such parents may be more concerned about their children's education and provide them additional academic assistance that retained students do not receive from their parents, but these potential family differences are usually ignored. The matching procedures are so inadequate in most of the citations provided by Holmes that few studies yield valid conclusions, regardless of how individual authors summarize their findings.

Although critics of grade retention extensively cite the Holmes (1989) review as definitive evidence against making students repeat a grade, they overlook Wilson's (1990, p.229) critique chiding Holmes for making "glaring mismeasurements and misinterpretations…." The major problem with most of the articles Holmes cited is their inability to adequately control for initial levels of academic achievement and ability between retained and promoted students. Even in most of the studies Holmes classified as better matched, retained and promoted students were dissimilar on initial levels of academic ability: Very few of these articles can be considered to be "tightly controlled." Consequently, conclusions from the Holmes meta-analysis about the impact of making students repeat a grade are not persuasive as many claim. One could argue that the research designs were better in the unexamined unpublished dissertations and masters' theses, but this is unlikely. Had the matching procedures or research methods been comparable to that used in the published articles, one would assume the graduate papers would have been published or

[9] Although Shepard et al. (1996) did not specify which of the studies listed in Holmes were the "six most tightly controlled," the average effect size from the Anfinson (1941), Archer (1967), Dobbs and Neville (1967), Schuyler and Matter (1983), Shepard and Smith (1987), and Wright (1979) studies is -0.24. The degree to which some of these studies control for initial differences is questionable (e.g., Dobbs and Neville, 1967; Shepard and Smith, 1987). In particular, Anfinson had no measure of student social and personal adjustment prior to retention.

[10] The five studies which attempted to match on achievement test and SES are listed here in order of study number Holmes provided. Following the study number is the reference to the study and the average computed effect size listed in Holmes (1989, p.27): #7=Dobbs and Neville (1967) -0.63; #17=Peterson et al. (1987) 0.76; #19=Schuyler and Matter (1983) -0.41; #20=Shepard and Smith (1987) -0.13; #25=Wright (1975) 0.35.

appeared in some other outlet. It is more probable that the problems of inadequate matching, lack of statistical controls, improper comparisons, and small sample size also pervade the dissertations and masters' theses cited by Holmes. Contrary to the assertions of many authorities, an unbiased reader should conclude that there is no overwhelming body of evidence in the Holmes (1989) meta-analysis to support the contention that grade retention is an ineffective or harmful educational remediation strategy. Many of the shortcomings found in the Holmes review are also prevalent in Jimerson's (2001) synthesis of more recent studies.

Jimerson Meta-Analysis

Jimerson (2001, hereafter Jimerson) published the most recent comprehensive review of grade retention studies. His meta-analysis contained 20 refereed articles, but only 18 examined academic achievement. Unlike Holmes, Jimerson made no effort to differentiate which of the studies in his synthesis utilised better-matching or statistical controls to account for initial differences between retained and promoted students. The indicators of research quality previously applied to the Holmes meta-analysis were also used to scrutinize Jimerson's review. First examined are the 16 studies in which the authors concluded that grade retention had no positive effect on student academic achievement. Similar to the review of the Holmes articles, studies cited by Jimerson are described in order of their rigor as determined by adequacy of the research design to control for potential student differences. Less rigorous studies are described first. However, an exception is sometimes made in the order of presentation. Articles based on the same data are jointly described regardless of the quality of research design. The two studies in which the authors favoured the practice of making students repeat a grade are examined last. Once again a detailed discussion of many articles is given. Although such a listing may seem monotonous, readers need to be aware of the unique features of these articles in order to form their own conclusions about the extent to which making students repeat a grade affects their academic performance. A summary of the major characteristics of the 18 studies from the Jimerson meta-analysis is presented in Table 2.

Comparable to the studies cited in the Holmes (1989) meta-analysis, most of the findings in the Jimerson survey are based on samples limited to only a small number of students. Of the 18 studies, 12 reported fewer than 60 pupils in either the retained or the nonretained control groups. Reynolds (1992) as well as Alexander et al. (1994) tracked the school performance of more than 200 academically challenged children. Meisels and Liaw (1993) were able to analyse the academic achievement of thousands of middle school students from a national representative sample of the U.S school population. With the exception of the Meisels and Liaw study, all of the other analyses were based on students from usually only one school district. The small sample sizes and restricted geographic locations of the studies limit the generalisability of the findings.

A major problem in many of these studies is again the lack of baseline data which would indicate if the promoted pupils were comparable to the retained student when they were held back in grade, an assumption underlying many of the articles Jimerson cited. None of the first four studies listed in Table 2 had indicators of student ability or outcome measures at the time of retention. Johnson, Merrell, and Stover (1990) compared 20 students retained in kindergarten or first grade from four public schools in the state of Washington with 17 pupils recommended for retention, but who were advanced to the next grade. Given that this was a retrospective study, there were no indicators of student ability or academic achievement prior to fourth grade. The outcome measures were fourth grade Metropolitan Achievement Test (MAT) scores. The absence of any statistically significant differences between the retained and promoted pupils recommended for repeating a grade, led Johnson et al. (1990) to conclude "… the use of early grade retention was not effective as an academic intervention…" (p.337), but the authors did not acknowledge that these academic differences may have existed prior to retention.

Table 2. Characteristics of Retention Studies in Jimerson (1989) Meta-Analysis

Author(s)	Quality of Controls	Outcome Pretest	Ability Pretest	Statistical Control for Outcome Pretest	Statistical Control for Ability Pretest	Type of Comparison	Grade Equivalent Units	Authors' Conclusion Effect of Retention
Johnson et al. (1990) (20-17-20)	Inadequate	No	No	No	No	Same-Grade	No	Negative[b]
Hagborg et al. (1991) (38-38)	Inadequate	No	No	No	No	Same-Grade	No	Negative
Meisels and Law. (1993) (3,203-13,420)	Inadequate	No	No	No	No	Same-Grade	Yes	Negative
Dennenbaum and Kulberg (1994) (25-28-17-25)	Inadequate	No	Yes[?]	No	No	Same-Grade	No	Negative
Ferguson (1991) (46-20)	Adequate	No	Yes	No	No	Same-Grade	No	Negative[b]
Ferguson and Mueller-Streib (1996) (33-14)	Inadequate	No	Yes[?]	No	No	Same-Grade	No	Negative
Phelps et al. (1992) (24-24)	Inadequate	Yes[a]	No	No	No	Same-Age	No	Negative[b]
McCombs-Thomas et al. (1992) (31-31)	Inadequate	Yes	No	No	No	Same-Age	No	Negative
Mantzicopoulos and Morrison (1992) (53-53)	Adequate[?]	Yes[?]	Yes[a]	No	No	Same-Grade Same-Age	No	Negative
Mantzicopoulos (1997) (25-15)	Adequate[?]	Yes[?]	Yes[a]	No	Yes	Same-Grade Same-Age	No	Negative
Rust and Wallace (1993) (60-60)	Adequate	Yes	No	No	No	Same-Age	No	Negative[b]
Jimerson et al. (1997) (32-50)	Adequate	Yes	Yes	No	No	Same-Grade	No	Negative[b]
Jimerson (1999) (20-23)	Inadequate	Yes	Yes[b]	No	No	Same-Grade	No	Negative
Reynolds (1992) (231-200-?)	Adequate	Yes	Yes	Yes	Yes	Same-Grade Same-Age	Yes	Negative
Reynolds and Bezruczko (1993) (251-1004?)	Adequate	Yes[a]	Yes[a]	Yes	Yes	Same-Age	No	Negative
McCoy and Reynolds (1999) (315-843)	Adequate	Yes[a]	Yes[a]	Yes	No	Same-Age	No	Negative
Pierson and Connell (1992) (74-69-35-60)	Adequate	Yes	Yes	No	No	Same-Age	No	Positive
Alexander et al. (1994) (242?-106?)	Adequate	Yes	No	Yes	No	Same-Grade Same-Age	No	Positive

[a] Retained and promoted students were significantly different on this variable at time of retention.
[b] Data indicate that retention had a positive effect on academic achievement.

Hagborg, Masella, Palladino, and Shepardson (1991) concurred that retention was ineffective after examining the academic performance of 38 high school students from a semi-rural New York school district who had been retained prior to eighth grade. The students' Comprehensive Test of Basic Skill scores (taken at the end of eighth grade), along with their Lorge-Thorndike Verbal and Non-verbal IQ scores (obtained at the end of seventh grade), were compared with the scores of students of the same sex who were in the same-track English class. On average, the 38 never-retained students had statistically significant higher scores. Such findings are not unexpected because the retained students were probably very different in initial levels of ability than exhibited by the promoted students who had never been retained. Without knowing how similar pupils were at the time of retention, any inference from this study is questionable.

The retention study cited in the Jimerson meta-analysis with the largest number of students is based on the 1988 National Education Longitudinal Study (NELS). Meisels and Liaw (1993) examined the grade retention histories (as reported by parents) of over 16,000 eighth grade students. More than 3,000 pupils had been retained once between kindergarten and grade eight. Results from reading and mathematics tests, prepared by the Educational Testing Service when the students were in eighth grade, along with self-reported grades were contrasted between the nonpromoted and promoted groups. Retained students were found to have significantly lower grades and standardised test results after controlling for gender, race, and family background characteristics. Meisels and Liaw (1993, p.76) concluded: "This study confirms that retention does not succeed in reducing the risks of later school failure." Such an inference was not substantiated, however, because there were no variables prior to eighth grade which measured earlier student academic ability.[11]

Three of the studies cited in Jimerson's summary were unable to match students on initial outcome measures, but instead attempted to match students on other indicators of student ability (Dennebaum and Kulberg, 1994; Ferguson, 1991; Ferguson and Mueller-Streib, 1996). Dennebaum and Kulberg (1994) tracked the academic progress of 95 elementary students from a Rhode Island school district. After examining school records of fourth and fifth graders, the authors created four groups of students. One group (n=25) had been retained in kindergarten before promotion to first grade. Twenty eight kindergartners were placed into a transition program before being promoted to first grade. A small number of kindergartners (n=17) had been recommended to repeat kindergarten or for placement in a transition room; instead they were advanced to first grade. A control group of 25 kindergartners were promoted directly to first grade. Reading, mathematics, language, and total battery scores from the Metropolitan Achievement Test, Sixth Edition (MAT-6) were the outcome measures for academic performance in first, second, and third grade. To determine if students were of comparable academic ability, mean scores from the Otis-Lennon Ability Test, Fifth Edition were examined at the end of first grade. The authors imply that the students in each of the four groups were comparable in ability because an analysis of variance revealed no statistically significant difference in the School Ability Test scores from the Otis-Lennon Ability Test. Students who had repeated kindergarten were observed to have lower fourth grade scores than the promoted students. Further, pupils placed in the pre-first transition grade also had lower fourth grade scores than the students in the two promoted groups. The authors asserted "the results indicate that retention actually hurt their achievement when compared to the children who were recommended for retention but went onto first grade anyway" (Dennebaum and Kulberg, 1994, p.11). They reasoned that retained students should perform significantly better than their unretained classmates who were a year younger. Although they do not present any data, the authors reported that no significant differences were found between the retained pupils and students in their classrooms who had not been required to repeat a year. The extra year spent repeating kindergarten or enrolling in the transition room was therefore viewed as a waste of time.

Though not evident, the research design is highly problematic because there were no indicators of student academic performance or cognitive ability when the 95 students were at the ***end*** of their kindergarten year. Outcome measures from the MAT-6 were initially observed at the end of first Grade, not kindergarten. No measure of student cognitive ability at the time of retention existed

[11] Critics of grade retention sometimes incorrectly summarize the nature of earlier studies. For example, McCoy and Reynolds (1999, p.276) report that Meisels and Liaw (1993) "adjusted for prior achievement" between the retained and promoted students in their article. This assertion is false because Meisels and Liaw (1993) clearly state that they were unable to control for academic achievement prior to or at the time of retention.

because the Otis-Lennon Test was not administered until the end of first Grade. Although the authors presumed the students were of comparable ability at the end of first Grade, the sample sizes of each group were probably not large enough to detect a statistically significant difference. Calculating effect sizes by dividing the difference between group means and the population standard deviation of 16 for the Student Ability Test, revealed that the retained kindergartners were, respectively, -0.50, -0.31, and -0.45 standard deviations behind the (a) transitioned, (b) recommended for retention/transition but promoted anyway, and (c) promoted pupils. The magnitudes of these standardized differences suggest that the retained kindergartners were of lower cognitive ability than students in the other three groups. The absence of an indicator of initial ability or an earlier measure of academic achievement as a statistical control nullifies the authors' conclusions.

The retained and socially promoted Wyoming elementary school students were more carefully matched on academic ability in Ferguson's (1991) study. Forty six kindergartners were required to attend a one-year transition first Grade while 20 pupils who had also been recommended for the transition readiness program were placed in first Grade, due to parental wishes or limited class size in the transition room. Students were matched on gender, age, and free/reduced lunch participation. Student development was measured by the Gesell School Readiness Test in the spring of the kindergarten year. The readiness scores of the two groups were comparable. Because no statistically significant mean differences between the promoted and retained pupils at the end of second Grade were found, the author concluded there was no academic advantage to placing students in a transition grade instead of promoting them to the next grade. Although some measurement specialists have questioned the validity of using the Gesell Tests to measure cognitive ability, the retained and control Wyoming students were probably of similar cognitive ability. Hence, the quality of controls is classified as adequate.[12]

Using the same data set, Ferguson reanalysed the Wyoming students when they were in fourth Grade and again found no statistically significant difference between the transition and promoted pupils (Ferguson and Mueller-Strieb, 1996). The authors concluded that making students repeat a grade did not increase later academic performance. Such an interpretation should be qualified because the sample sizes of 33 transitioned students and 14 promoted pupils were even smaller than in the previous analyses of second Graders, which further reduces the likelihood of finding a statistically significant difference. Unlike the prior study, the authors did not indicate if the transitioned and promoted students had similar Gesell Readiness scores. Given the unknown degree of cognitive comparability between the retained and promoted groups, this article is listed as having an inadequate research design. The authors could have strengthened their conclusions by using the initial Gesell score as a covariate to predict the fourth Grade achievement scores.

The remaining retention studies shown in Table 2 were able to obtain initial measures of student academic performance. However, the research designs in Phelps, Dowdell, Rizzo, Ehrlich, and Wilczenski (1992) and McCombs-Thomas et al. (1992) were classified as inadequate because the authors made no effort to control for existing differences between promoted and nonpromoted students at the time of retention. Phelps et al. (1992) attempted to equalize initial differences by matching on individual characteristics across three student groups examined in a blue-collar suburban Buffalo, New York school district. The first set of students, classified by teachers as

---

[12] A later analysis of these data (Ferguson and Mueller-Strieb 1996) revealed that other measures of early development were available (e.g., the Brigance K and 1 Screening Test as well as the Metropolitan Readiness Test). It would have been useful to ascertain if the students placed in the transition room and their promoted peers differed on these measures which may be better measures of student ability than the Gesell School Readiness Tests.

"slow learners" had been placed in a pre-grade transition room instead of being promoted directly to the next grade. Of the 22 pupils in this group, 16 were placed in a pre-first grade transition room; five students were required to attend a pre-second grade transition room; one student was placed in a pre-third grade room. Another 24 students in grades two through four were viewed by their teachers as developmentally immature. These pupils were required to repeat their present grade. A control group of 24 students were matched on gender, free lunch status, IQ, and current grade placement with the transitioned and retained pupils. The students in the control group were current classmates of the transitioned and retained children who had never been retained. Outcome measures were obtained five to ten years after retention. There were two general findings. First, there were no statistically significant differences in reading scores across the three groups. Second, the control students had significantly higher mathematics scores than the transitioned and retained students. The authors concluded: "Results suggested that neither retention nor transition placement resolved the academic deficiencies of these children" (p.112).

However, this interpretation is questionable because the authors did not adequately control for initial differences in academic achievement. Reading and mathematics scores from the California Achievement Test (CAT) were obtained at the end of first Grade. Instead of reporting ability scores at the end of kindergarten when students were retained, the pretest scores for the 16 transitioned kindergartners (73 percent of the group) were based on their end-of-year pre-first transition grade, one year after their placement. The absence of end-of-year kindergarten test results limits the ability to assess comparability across the three groups. Given that this was a retrospective study, the CAT scores reported five to ten years after being transitioned or retained were obtained from school records, when the students were in Grades seven through nine. While the IQ scores of the three groups were similar, the grade level of initial measurement was not reported. The study does not indicate if IQ scores were obtained before or after retention. A more meaningful summary of the effect of the various educational practices could have been obtained had the authors performed an analysis of covariance using the end of first Grade CAT scores as a covariate to estimate group differences in the later grades.

The degree to which retained and promoted students were similar at the time of retention was also uncertain in analyses based on a small number of pupils in a rural school district (McCombs-Thomas et al., 1992). Students who were retained in kindergarten or first Grade (n=31) were matched on race, gender, and grade point average, at the time of retention, with 31 nonretained students. Student grade point averages when the children were in Grades two through five were the outcome measures. No statistically significant differences in grade point averages were observed in any of the grades. However, separate analyses by race revealed the retained white students had significantly lower GPAs in third and fifth Grade than their promoted white counterparts. The authors did not specify if the initial grades were the same between the retained and nonretained white elementary pupils at the time of retention. McCombs-Thomas et al. (p.347) even acknowledged that the two groups of students may not have been comparable because teachers described the retained pupils as progressing much more slowly in their coursework than the promoted students. The two groups had similar initial GPAs because teachers graded on the work students had completed; however, the retained students were required to repeat either kindergarten or first Grade because they could not keep up with their promoted classmates. The authors then ignored the problem of comparability by stating that the students could be matched only on these limited numbers of variables. They imply that inability to obtain similar students for comparative purposes should be overlooked when assessing the impact of educational programs.

Nine of the 18 studies cited in Jimerson's review met the criteria required to be classified as having adequate controls, although various dimensions of the studies were still problematic. A sample of 53 retained Marin County, California kindergartners were matched with 53 kindergarten same-grade peers who had not been required to repeat the grade (Mantzicopoulos

and Morrison, 1992). Pupils were matched on age, socioeconomic status, and academic achievement, calculated from the Stanford Achievement Test (SAT) or the California Test of Basic Skills (CTBS). Further, a screening instrument was used to determine if the pupils were at-risk for reading failure. The two groups initially differed by 0.2 of a standard deviation in achievement scores. The probability of a retained student being at-risk of failing reading was 0.71 whereas the likelihood of have reading difficulty among the promoted was 0.51. After repeating kindergarten, the retained students' second set of kindergarten reading and mathematics achievement scores were significantly higher than the initial kindergarten scores of the promoted pupils. However, because there were no statistically significant lasting gains from the first and second Grade comparisons, the authors presumed that making students repeat a grade was not beneficial.

Two issues make interpretation of the findings difficult; the first pertaining to measurement procedures. The retained kindergartners were pooled from two different school districts in two different academic years. Given that one district used the Stanford Achievement Test to measure academic achievement and the other district the California Test of Basic Skills, the authors converted the initial responses from the two different tests to standard scores in an attempt to make them comparable. It is unclear as to how the *z* scores were computed. Combining z scores based on national norms from two different tests is problematic because the norms were not based on the same populations. Only if all the retained and promoted students took both the SAT and CTBS would *z* scores be appropriate to create a combined index of academic achievement, but these children took different exams. A more problematic issue is that the authors ignored the significantly higher observed levels of immaturity among the retained students (Mantzicopoulos and Morrison, 1992, p.193). The promoted and retained students were not comparably matched in abilities that could affect academic achievement. Had the study utilized initial levels of academic performance and behavioural characteristics as covariates, the results could have been more easily interpreted because plausible rival explanations of earlier group differences would have been eliminated.

In a later article Mantzicopoulos (1997) attempted to gauge the impact of grade retention among the students with high inattention scores. Twenty five children with high inattention (that is, low maturity) scores were contrasted with 15 control students from the initial study who also had high inattention ratings (Mantzicopoulos and Morrison, 1992). Math and reading achievement *z* scores derived from the two different tests from the earlier study were again the outcome measures of achievement. Children's attention problem scores were used as a covariate. Findings based on both same-grade and same-age comparisons indicated that the retained students outperformed the nonretained pupils on the mathematics test, but no advantage occurred on the reading test. As was the case with the earlier study, the initial achievement measures were not used as covariates. The same-grade comparisons were made only in first and second Grades. A more appropriate time frame would have been to use outcome scores as was done in the first paper based on the total number of measurement periods, that is, the contrast between retained and promoted children at the end of the second year of kindergarten, first Grade, and then second Grade. The initial paper revealed that the positive impact of holding children back occurred during the year of retention, not the year after. Mantzicopoulos's findings may be biased because initial scores from the end of the first year of kindergarten were ignored.

A similar analysis was conducted on Tennessee students. Rust and Wallace (1993) matched 60 students, who had either been retained in kindergarten or placed in a transition room, with 60 low-achieving pupils who were promoted to first Grade. Students were matched on race, gender, free lunch status, and classroom grades prior to retention. Outcome measures were student grade point averages and national standardised tests (that is, the California Achievement Test and various forms of the Stanford Achievement Tests). The authors did not explicitly state if same-age or

same-grade comparisons were made, however, readers were given the impression that nonpromoted and promoted pupils were evaluated when the promoted pupils were one year ahead of the retainees (that is, same-age comparisons). Although both groups of students had similar grades at the end of the kindergarten year, the retained/transitioned pupils scored significantly lower on the kindergarten achievement test. The average nationally-normed test score of those students who were retained at the end of the school year was 0.27 standard deviations below that of the promoted students. At the end of the next year pupils were retested, after the retained/transitioned students were held back and the nonretained students were at the end of the first Grade year. The retained students obtained significantly higher scores on the national standardised examination than the promoted; the effect size was 0.27 in favour of those held back a year. However, no significant differences in test scores were observed over the next two years. Classroom grades for the retained/transitioned children were also higher in the second and third years of the study, but the differences with the promoted children were not statistically significant. The authors argued against holding students back a year because "This study found weak evidence that retention may benefit children" (Rust and Wallace, 1993, p.165). Given that comparable nonretained low-achieving students were performing successfully in the next grades, the authors did not recommend making students repeat a grade. As occurred in the previous studies, the authors did not statistically control for initial differences in test performance. Using the kindergarten test results as a covariate to help control for initial differences in academic ability between the promoted and nonpromoted children may have resulted in a more accurate assessment of the impact of grade retention.

Jimerson's review of retention studies also cites his research which tracks the academic success of a small number of Minnesota children through their teen and early adult years (Jimerson, Carlson, Rotert, Egeland, and Sroufe, 1997). Thirty two students who had been retained between kindergarten and third Grade were initially matched with 50 equally low-achieving children in the same grades who had been promoted. Nonretained students who fell in the bottom quartile of the Peabody Individual Achievement Test (PIAT) in first, second, or third Grade were selected for comparison purposes. Teacher ratings of kindergarten students were used to obtain a comparison group for the earliest retainees because the PIAT was not administered at the end of kindergarten. Children required to repeat kindergarten were given the PIAT at the end of their second year in the grade. Retained students and their matched counterparts were found to have statistically similar PIAT scores as well as similar levels of intelligence, as measured by the Wechsler Preschool and Primary Scales of Intelligence (WPPSI). Relative to the low-achieving promoted control group, the retained students evidenced greater behavioural and emotional-health problems when in kindergarten.

The first set of analyses assessed the short-term effects of holding students back in grade by examining academic outcome indicators after the year of retention. Measures of academic performance among the kindergarten retainees, however, were taken at the completion of first grade. Given that only nine retained kindergarteners were contrasted with 15 matched pupils, it is to be expected that no statistically significant differences existed between the two groups. In addition to the small numbers of students for comparative purposes, no data existed to measure the degree of academic achievement when both the retained and promoted students would have been completing kindergarten. The retained pupils could have had significantly lower levels of achievement at the end of kindergarten than their same-grade classmates. A stronger case for the ineffectiveness of kindergarten retention could have been made by statistically controlling for the Wechsler IQ measure (WPPSI), which was taken when the students were age five, prior to kindergarten retention. Although the means of the two groups were not statistically different, the retained children were 0.2 of a standard deviation behind the low-achieving promoted students when tested in kindergarten.

A better statistical adjustment was possible when examining the impact of first or second Grade retention because the PIAT achievement score at the time of retention was entered as a covariate to predict later academic achievement. The retained students significantly outperformed their low-achieving promoted classmates in mathematics, but not in reading. Again, small samples reduced the power of statistical tests as only 16 retained pupils were matched with 28 promoted students. Additional comparisons between the retained and matched students were made at the end of sixth Grade and when the students were age 16. At both times, no significant mean differences in academic achievement emerged.

Jimerson (1999) also reported on the academic achievement when study participants were in Grade eleven and later. The number of study participants with complete data decreased considerably over time. Only 20 students who had been retained before fourth Grade were available to be compared with 23 matched controls. Because many of the student's family, social, and earlier levels of academic achievement were presumed to be similar, Jimerson made no attempt to control for these groups differences. He found that, when compared to their promoted low-achieving counterparts, the retained students had significantly lower levels of academic adjustment when in 11th Grade, were more likely to have dropped out of high school by age 19, and have a lower probability of completing high school by age 20. Jimerson concluded retention led to "poorer" academic outcomes.

The first study (Jimerson et al., 1997) attempted to statistically control for baseline measures of academic performance. Similar statistical controls would have resulted in greater credibility to his conclusions in the second study (Jimerson, 1999). The earlier study indicated that, when in the elementary grades, students were dissimilar on important characteristics that could affect later academic achievement. For example, the measure of student intelligence (Wechsler Intelligence Scale for Children-Revised), taken when the children were in third Grade, was found to be significantly different between the retained and promoted students. The retainees were 0.37 of a standard deviation below the mean score of their matched counterparts prior to high school (Jimerson, 1999, p.257). Another indication that the promoted and nonpromoted students were dissimilar was that the retained students had missed a significantly greater percentage of school days than the other low-achieving children (p.257). State laws often mandate that students repeat a grade if they are absent a certain number of days. Although Jimerson noted that various reasons can be used to retain students, he made no attempt to determine if differences existed in the intelligence or academic achievement of students who were retained, probably because of absences, with those who were promoted. For these reasons, Jimerson's (1999) later article is labelled inadequate.

These initial differences between the two groups of students raise the question as to whether the later achievement differences between the retainees and the promoted can actually be attributed to making students repeat a grade in elementary school. The high school measure of academic adjustment Jimerson created is problematic because it is a composite index of three separate indicators: high school achievement (grade point average and ratio of high school credits to number of years in high school), behavioural problems, and attendance. However, the retained and socially promoted peers differed significantly in the elementary grades on the behavioural problems and attendance measures. Even with small sample sizes, it would have been appropriate to incorporate the earlier elementary school indicators of these behaviours in separate regression equations so as to help rule out the possibility that differences observed in high school were not due to initial behavioural differences between the socially promoted and retained students. Similarly, the PIAT first grade measure of academic achievement and the third grade WISC-R could be used as covariates in the analyses to predict later academic achievement. Like many other analyses of the impact of grade retention on student academic performance, both the Jimerson

studies could have better controlled for initial differences between the promoted and nonpromoted children.

One of the few other analyses listed in Jimerson's (2001) review which utilised regression techniques to control for initial differences between retained and promoted pupils followed the academic achievement of inner city African American children attending Chicago public schools (Reynolds, 1992). The outcome measures were Iowa Test of Basic Skills (ITBS) reading and mathematics scores, along with teacher evaluations of the students' academic competence. Two hundred and thirty one elementary school students who had been required to repeat any grade between kindergarten and third Grade were contrasted with 1000 "regular" children who had never been retained. More important was the availability of a control group consisting of 200 similar low-achieving students who had not been retained. In addition to establishing a control group of students with matched ability, social, and, psychological characteristics similar to the retained children, Reynolds statistically adjusted for possible differences between the two low-achieving groups. For example, Grade one ITBS scores were included as covariates to control for possible differences in academic ability. Same-age comparisons were made in which the promoted pupils were in Grade four while the retained students were still in Grade three. Consistent with findings from other studies, the academic achievement of retained pupils significantly lagged behind the normal students who had never been required to repeat a grade. More importantly the retained students also exhibited lower scores in reading and mathematics than the low-achieving control group of promoted children. The large negative effect sizes for the retained students led Reynolds (1992, p.117) to conclude that the effects of grade retention were negative and harmful or, at best, negligible.

Jimerson also cited two other studies based on the same data. Reynolds reported the results from an almost identical analysis, except only reading scores were the outcome measure (Reynolds and Bezruczko, 1993). Whereas the earlier 1992 study examined ITBS grade-equivalent scores, the outcome measures in the 1993 article were transformed into logit values derived from a one-parameter item response theory method. Unlike the previous study, the control group consisted of all students who had never been required to repeat a grade. The control group of comparable low-achieving students described in the 1992 analyses was not utilized. Grade retention was again found to have a negative net effect on reading performance based on the transformed ITBS reading scores. Academic achievement levels of the same retained and promoted Chicago elementary school students were also assessed later when the adolescents were age 14 (McCoy and Reynolds, 1999). Both same-age and same-grade comparisons in this last article revealed students who had been retained at some point between Grades one and seven significantly underperformed in both reading and mathematics when compared to the continually promoted adolescents.

The Reynolds analyses are superior in several ways to most other retention studies. The large numbers of retained and promoted students enable detection of significant differences unlikely to be found in studies with small samples. A major positive feature is that, rather than only relying on a matching procedure, Reynolds was able to statistically control for many determinants of academic achievement prior to or near the time of retention. Most important is the availability of similar low-achieving pupils reported in the first study (Reynolds, 1992) which allowed for the creation of a control group more comparable to the retained pupils. All of these features result in a more accurate assessment of the impact of grade retention on academic performance. Nonetheless, the Reynolds papers possess features that raise questions about the certainty of the conclusions.

The higher academic achievement of the promoted pupils is not unexpected, even after statistically controlling for initial differences in family background, cognitive readiness, and other psychological traits, because the total promoted group of pupils (n=1000) evidenced significantly

higher baseline scores. Jackson (1975) argued that such a "comparison is biased toward indicating that grade promotion has more benefits than grade retention because it compares retained students who usually are not having as severe difficulties, as evidenced by the fact that they have not been retained in grade" (p.619). Campbell and Kenny (1999) point out that if the treated group (for example, the retained students) has lower scores than the control group (for example, promoted pupils), an analysis of covariance will still likely under adjust for initial differences because of the "failure to measure and control for the variable that is used to assign persons to treatment groups" (Campbell and Kenny, 1999, pp.75-76). Within the context of the effects of grade retention, the inability to properly specify those variables that actually caused the student to be retained yields coefficients that imply holding students back a grade is more negative than is likely to be the case. A related shortcoming is the absence of the comparable 200 low-achieving children in the control group reported in the initial study (Reynolds, 1992) which was not utilised in either of the latter two papers (that is, Reynolds and Bezruczko, 1993; McCoy and Reynolds, 1999). Given the greater equivalency in ability measures of these matched students with the retained pupils, findings indicating higher academic achievement of these students compared to the retained would have further strengthened the argument against retention.

Closer examination of the reported differences in academic performance between the retained and comparable promoted students in the initial article (Reynolds, 1992) raises additional concerns about the findings. Reynolds analysed ITBS scores which had been converted to grade-equivalent scores. Although Reynolds (1992, pp.118-119) noted that grade level scores were sometimes misinterpreted, he assumed the psychometric properties of grade level scores were adequate for comparative purposes. However, he overlooks the criticism that grade level scores do not allow adequate measurement of change over time between groups.[13]

The type of group comparison Reynolds (1992) used may be unclear to the casual reader. For simplicity Reynolds used the term *same-grade* to describe the nature of the comparison between the retained and promoted students (p.104). However, the term *same-grade* is commonly used to denote a comparison when pupils are in the same grade. The appropriate definition for the comparisons Reynolds performed is *same-age* because outcome measures were based on scores measured in the same year when students were in different grades (in this study Grade three for the retained pupils and fourth Grade for the promoted students). Because nonretained students cover an additional year of instruction in a later grade, same-age comparisons usually result in the promoted students demonstrating higher levels of achievement than the retained. Having been exposed to a more advanced curriculum for about eight month would help explain why the ITBS reading and mathematics grade-equivalent scores of the matched promoted group were about 0.8 of a year ahead of the retained students; that is when the retained students were in third Grade and their promoted counterparts were in Grade four.

Had Reynolds (1992) compared the academic performance of the 200 pupils in the matched promoted group students when both sets of students were in third Grade, (instead of different grades the same year), the negative results of making students repeat a grade may not have been as pronounced. The matched promoted students would not have had an additional year of exposure to the fourth Grade curriculum which probably raised their grade-equivalent test scores. Although not a same-grade comparison as is traditionally defined in retention research, Reynolds (1992,

---

[13] Reynolds (1992, p. 118) attempted to justify the use of grade-equivalent scores by noting that their correlation with logit values is 0.98; however, readers are not informed about the nature of the correlations between grade-equivalent and logit scores over time for the three groups of students examined. Equal-interval logit scores appear in his later two articles (Reynolds and Bezruczko, 1993; McCoy and Reynolds 1989).

p.113) compared grade-equivalent test scores for the retained students when they were in Grade three with those of "matched-grade peers." These third Grade students were likely to be a year younger than the retained students because these matched-grade peers had never been required to repeat a grade. Although the effect size for reading was -0.12 and 0.07 for mathematics, the third Grade test results did not significantly differ. Reynolds (1992, p.113) therefore concluded that "retention is unrelated to academic achievement" because he assumed the retained students should have had higher levels of academic achievement since they were a year older. Such an assumption seems questionable because the most advanced curriculum the retained students had been introduced to was the same as that of their third Grade classmates. Although the effect sizes between the two groups of children were negligible and statistically insignificant, it is not unreasonable to speculate that the initial first Grade ITBS scores of the retained third Graders were significantly below those of their nonretained third Grade classmates, otherwise they too should have been held back a year. Were the matched-grade peers to have higher scores in first Grade than the retained students, one could reasonably argue that making the low-performing pupils repeat a grade helped bring them up to the level of their classmates who were continually promoted to third Grade. Reynolds provides no information in his article to assess the reasonableness of this hypothesis.

Only two of the studies examined in the Jimerson (2001) meta-analysis concluded that requiring low-performing students to repeat a grade improved academic achievement. Pierson and Connell (1992) compared the academic performance of 74 upstate New York students retained in grades one through four (most were required to repeat first Grade) with 69 promoted same-grade peers matched on Otis-Lennon Mental Ability IQ scores, sex, and grade when both groups of students were in their current grade. An additional comparison group consisted of 35 students who had been placed in the next grade, that is, the pupils should not have been promoted because of grades or the teacher's recommendation. These children were matched with the retained pupils on the basis of similar grade point average, sex, and grade. Grade point averages were comparable between the retained and socially promoted pupils at the time of retention (Pierson and Connell, 1992, p.303). The overall outcome measure of academic performance was based on the (a) mean value of final marks for all academic subjects averaged from Grades three through six and (b) the mean reading and mathematics results from national standardized percentiles from Grades two through six (that is, the grades after retention). These two indicators of academic performance were averaged to measure overall achievement.

The retained pupils achieved academic outcome scores similar to their peers with comparable IQs. More interesting was the finding that the retained students performed significantly better on the global indicator of academic achievement than the socially promoted students. The mean academic achievement score for the retained students was 0.56 standard deviations above that of the socially promoted. Consequently, the authors argued that, while not a "cure-all" for below grade-level academic performance, "the findings support the use of retention as a potentially effective remediation for academic difficulty in the early elementary grades" (Pierson and Connell 1992, p.306). A limitation of the study, however, is that no annual academic achievement data were presented, only the aggregate of grades and standardized test scores over the period of study. It is not possible to determine whether the positive impact of retention occurred only after the year of being held back, or whether the difference in achievement scores persisted over time. A unique aspect of this study is that no other articles combined annual performance indicators into a single score; instead, other studies reported yearly achievement measures.

The second study concluding in favour of retention was a prospective study of a random sample of 800 elementary students in the Baltimore City Public Schools (Alexander et al., 1994). Study participants were followed from the fall of 1982 when the students were entering first Grade through the spring of 1990. Over 300 students had been required to repeat a grade at least

once. Given the many different kinds of outcome measures analysed, it is difficult to determine the exact number of students who were retained or promoted with usable data. Academic achievement was measured by use of the California Achievement Tests (both reading and mathematics) and course grades in reading and mathematics. There were 242 students who repeated either first, second, or third Grade. The major control group of interest was 106 never-retained poor-performing children, although comparisons were also made with regularly promoted students. Students were examined by grade of retention. The actual number of students analysed varied because pairwise deletion of data was used to maximize the number of observations. After following students from first Grade through eighth Grade, the authors concluded that making students repeat a grade helped boost their academic ability so that they performed at levels closer to those of the regularly promoted students, although the nonpromoted children never caught up with those students who experienced no academic difficulties.

Before evaluating Jimerson's overall conclusion regarding the meaning of all these studies reviewed, it is worthwhile to point out several types of errors which occurred in his meta-analysis. One type of error Dunkin (1996) identified is "listing different reports from the same project as providing additional confirmation of the same finding" (p.91). This "double counting" appeared several times in the Jimerson (2001) meta-analysis. Two studies were based on a small sample of Wyoming elementary school children (Ferguson, 1991; Ferguson and Mueller-Streib, 1996). Students sampled from Marin County appeared in two papers (Mantzicopoulos and Morrison, 1992; Mantzicopoulos, 1997). Jimerson also cited his two analyses which used participants from the same Minnesota sample (Jimerson et al., 1997; Jimerson, 1999). Similarly, findings from the three papers by Reynolds and his colleagues (Reynolds, 1992; Reynolds and Bezruczko, 1993; McCoy and Reynolds, 1999), which utilized the same sample of low-income African American children from Chicago, were counted as if they were from different data sets. In all of these articles, the authors concluded that grade retention diminished student academic achievement. Following the same students over time will likely yield the same negative comparisons between retained and promoted students favouring the latter. The result of treating identical data sets as if they were independent results in an inflated number of negative outcomes.

Syntheses of literature are sometimes in error because an author incorrectly describes the methodology or context of a study. Dunkin (1996, p.90) refers to this kind of incorrect statement as "erroneous detailing." Jimerson (2001) made this error several times in his meta-analysis when listing the characteristics of students in the control groups. The first instance occurred when describing the Hagborg et al. (1991) study. Jimerson listed the article as controlling for academic ability because the authors presumed students in the same high school classes were of equal ability. In actuality, there were no prior indicators of student ability. Jimerson committed another "detailing error" when describing the Dennebaum and Kullberg (1994) study because the measure of ability used to match students was not made at the time of retention but after. His description of the McCombs-Thomas et al. (1992) analyses was partially incorrect because, contrary to the information listed in Jimerson's Table 1, IQ measures were not mentioned as a basis of matching students. The authors of these three studies concluded that retention was ineffective. However, these errors in study details may incorrectly lead readers into thinking the retained and promoted students were more similar than was the case. Another detailing error arose when describing the comparison groups in the Pierson and Connell (1992) study which concluded in favour of retention. Jimerson correctly described one comparison group (n=69) which did not have data indicting initial academic ability. However, Jimerson ignored a second comparison group (n=35) of socially promoted children matched on similar grade point averages with the retained students. These socially promoted pupils evidenced significantly lower grade point averages than the retained at the end of the study period. Given the large number of articles and variables reviewed, it is quite easy to incorrectly list the detailed characteristics of an article. Nonetheless, these kinds

of errors may shade the overall conclusions to be made when aggregating the findings from individual studies.

While Jimerson may not have committed a detailing error per se when interpreting the study of Baltimore elementary students, which concluded in favour of retention, a different interpretation of the findings is possible. The Alexander et al. (1994) Baltimore study is particularly important in Jimerson's review because 47 per cent of the negative findings reported in his meta-analysis were derived from this analysis. Jimerson summarized both the same-age and same-grade comparisons Alexander and his colleagues calculated. A count of the same-age comparisons resulted in 23 significant negative effects, and seven contrasts which did not differ. These comparisons were based on mean differences between the retained and a group of poor-performing children, after adjusting for initial social-demographic differences and an indictor of student test performance prior to retention between the two groups. With respect to same-grade differences, Jimerson reported a total of 21 statistically significant negative effects, 28 nonsignificant coefficients, and only one significant positive effect for grade retention. These comparisons do not include the results measured at the end of the retention year.[14] In addition, Jimerson ignored a set of more refined comparisons which also controlled for students who were retained more than once and students classified in special education. Under these different circumstances the summary of findings changed considerably. A recount of the same-grade comparisons pertaining to test scores and classroom grades revealed only eight significant negative effects and 49 insignificant differences. Three contrasts favoured the retained students. A reexamination of the Alexander et al. (1994) study suggests that making students repeat a grade is more positive than Jimerson reported. This conclusion, however, is based on a different set of comparative criteria which controls for more factors. Rather than count the number of negative, insignificant, or positive results from specific contrasts between the retained and poor-performing control group, Alexander and his colleagues argue that it is more important to examine the overall pattern of findings. The more worthwhile question from their perspective is whether retention helped raise the academic performance of children required to repeat a grade. The 12 same-grade comparisons presented in Alexander et al. (1994) which control for double retention and special education characteristics reveals that prior to being held back, the retainees were significantly below the comparisons group. But after repeating the grade, the test scores and grades of the retained students were equal to those of the promoted poor-performing children. Consequently, Alexander et al. (1994) contend their findings demonstrate that retention can help enhance academic achievement because "for retainees just to be holding their own may be an accomplishment" (p.22).

Sixteen of the 18 studies listed in Jimerson's meta-analysis did not favour making students repeat a grade to improve their educational performance. As was also observed in the Holmes (1989) review, authors of individual articles overlooked findings which contradicted their overall conclusions. Partial results from four of the studies in the Jimerson meta-analysis could be interpreted to suggest that grade retention is an effective strategy to help raise student achievement. Because no statistically significant differences were found between retained and promoted students, various authors concluded holding students back a year in school was ineffective. Had the authors computed effect sizes between the two groups, they may have

---

[14] Professor S. Jimerson (personal communication, June 22, 2004) was kind enough to provide me information describing how he derived the numbers summarizing the number of statistically insignificant, negative, and positive mean differences between the retained and poor-performing promoted students in the Alexander et al. (1994) study. The aggregate counts in Jimerson's (2001, p. 426) summary were derived from the information Alexander and his colleagues presented in Chapters 5 and 6 of their book. The number of insignificant comparisons is actually 38 instead of the 28 shown in Jimerson's Table 1.

changed their conclusions. For example, the investigation of Washington state students (Johnson et al., 1990) failed to mention that fourth-Grade total math and math calculation scores were higher among the retained. If one computes effect sizes for these two subscales from the Metropolitan Achievement Test, the respective values are 0.44 and 0.42, which imply that retention helped increase mathematics performance. Similarly, Ferguson's (1991) analysis of Wyoming pupils showed the mean values from the SRA Achievement Test taken in second Grade of children put in the readiness transition room were numerically larger than those of the promoted kindergartners. Computing effect sizes for the mean differences yielded values in favour of the retained pupils. The effect sizes for each tested area were 1.17 (language), 0.86 (math), 0.28 (reading), and 0.88 (total score). These findings imply that placement in the transition first Grade was helpful, particularly in three of the four areas tested. Had Jimerson et al. (1997) calculated effect sizes between retained and nonretained students for certain outcome measures, they would have found that the children who repeated a grade outperformed those who were promoted. The effect size between the two groups for the total score of the PIAT at the end of first Grade was 0.38 standard deviation units in favour of those retained in kindergarten. Among the first Grade and second Grade retainees, the effect sizes for the PIAT mathematics test were substantial for both the unadjusted means (1.21) and the adjusted means (1.07). With respect to the PIAT total score, the retained first and second Graders outperformed the promoted students by 0.74 standard deviations on the observed means and 0.64 standard deviations for the adjusted means. If the authors of these three studies assumed effect sizes of 0.20 or higher indicated "educationally meaningful" differences, as Reynolds (1992, p.107) suggested, they could have concluded that holding low-performing students back a year provided an educational benefit. None of these authors calculated effect sizes because there were no statistically significant differences between the retained and promoted students. However, the number of observations in each comparison group was so small that even substantively large calculated effect sizes were deemed insignificant. Group means were not statistically significant because of the small sample sizes with insufficient power to reject null hypotheses, even when the levels of academic achievement were substantially different in standard deviation units.

The magnitudes of the effect sizes suggest that the authors' conclusions regarding their findings may be based more on their subjective interpretations rather than the specific results from the data. Does an insignificant difference between retained and promoted students indicate that retention is effective? The interpretation partially depends on what authors assume is needed as proof of effectiveness. The study of New York students by Phelps et al. (1992) revealed that the pre-retention reading and mathematics means of the transitioned and retained students were significantly lower than that of the control group. By the end of the period under investigation, however, mean reading values were comparable across the groups. This similarity could be interpreted to indicate that making students repeat a grade helped reduce the initial differences among the three groups. That is, the lower-performing pupils who were held back a year caught up with their promoted classmates. But Phelps and her colleagues assumed that the scores of the retained students should surpass that of the nonretained children. The absence of standard deviations in this study prevents readers from calculating effect sizes between the transitioned, retained, and matched control group. However, an examination of the means shows that the students placed in the transition room raised their reading scores by nine percent while the retained students increased their reading scores by over seven percent. Conversely the reading scores of the promoted control group decreased by almost five percent. Regardless of the authors' conclusion, these findings reveal that holding students back a year at least enables them to catch up with their promoted classmates and not fall further behind.

As has been stressed throughout this discussion, the inability to control for initial levels of academic achievement between retained and promoted groups can lead to only ambiguous

conclusions about the impact of holding students back. Even though several of the authors recognized this limitation of their analyses, they nonetheless concluded that low-performing students should not be required to repeat a grade. The Phelps et al. (1992) study aptly illustrates this type of reasoning. The authors' acknowledge that (a) only randomization studies can prove causation and (b) matching on available variable can not control for all sources of variability (Phelps et al., 1992, pp.121-122). But they made no attempt to statistically control for initial differences between retained and promoted students and negatively assessed the impact of transition room placement and grade retention.

Meisels and Liaw (1993) reached a similar negative conclusion, although they recognised the difficulty in interpreting their findings: "it is possible that the retained students who showed less optimal academic performance in eighth Grade were academically less able or had problems that predated retention" (p.75). Nonetheless, the authors rejected this alternative interpretation because (a) their findings were similar to those summarised in the Holmes (1989) meta-analysis and (b) they analysed a large national representative sample of students. As was previously mentioned, a detailed examination of the published retention studies in the Holmes (1989) summary does not yield a consistent pattern of negative effects because most of the research designs do not incorporate baseline measures of student ability. The fact that most of the retention studies lack an earlier indicator of student academic performance does not justify the Meisels and Liaw (1993) conclusion. Hagborg et al. (1991) likewise concluded that making students repeat a grade is probably ineffective because the retained students did not perform as well as their continually promoted classmates. The lack of earlier data measuring student aptitude, which could indicate how far behind the retainees fell relative to their same-grade peers, does not allow Hagborg and his coauthors to infer that retention was not helpful. Rather than recognize the limited nature of their study, the authors relied on the questionable conclusion of an earlier meta-analysis to support their position: "However, given the doubtful benefits of retention (Holmes and Matthews, 1984), it is possible that the educational needs of retained students were not adequately addressed, and they were left behind their classmates" (Hagborg et al., 1991, p.315). By making no effort to evaluate the validity of many of the conclusions offered in the reviewed studies, Jimerson (2001) committed the "Nonrecognition of faulty author conclusions" discussed by Dunkin (1996, p.91).

Overlooked and More Recent Research

Before summarising the Holmes and Jimerson meta-analyses, it is worthwhile to examine several recent published articles. A summary of the research characteristics of these studies is presented in Table 3. One investigation analysed the academic achievement of largely Hispanic elementary children in south central California (Cosden, Zimmer, Reyes, and Gutierrez, 1995). Kindergarten students who had been required to repeat the grade were matched on birth month, gender, ethnicity/home language with an equal number of classmates at the end of first Grade who had never been retained. Low-performing kindergartners who had been "advanced" (that is, placed) in first Grade in spite of observed academic difficulties were also matched using the same variables with an equal number of never-retained classmates. Achievement was measured at the end of the first Grade year by the Stanford Achievement Test for English speaking students and the Aprenda among Spanish speaking children. Preliminary MANOVAs on reading, language, mathematics, and the basic battery indicated that the English speaking retained (n=17) and advanced kindergartners (n=35) had lower scores than their never-retained classmates. Mean first Grade scores on the Aprenda were comparable between the retained Spanish speaking children (n=19) and their classmates, but the reading scores of the Latino students who had been advanced into first Grade were significantly below their matched promoted Spanish speaking first Grade peers. Multiple regression analyses which combined all of the students revealed that only the low-achieving advanced children (both the English and Spanish tested students) had significantly

lower scores than their controls. Cosden, et al. (1995) concluded "that neither intervention [retention or advancement] had a positive impact on the students' achievement" (p.137). However, the authors correctly acknowledged that the study could not assess the extent to which retention or advancement influenced academic results because of uncontrolled factors, such as parental education and quality of relationships with the schools. Like most of the studies cited in the retention meta-analyses, no baseline measures of kindergarten performance were available as control variables. The academic skills of the retained and advanced students were probably below their first Grade classmates even in kindergarten.

Findings from two other studies with better controls for initial ability suggest that making students repeat a grade may increase academic achievement. The first study investigated the effect of making students in a west central Florida semirural county school district repeat a grade. Pomplun (1988) examined the change in NCE scores on the Comprehensive Tests of Basic Skills among retained students, borderline pupils who had been placed in the next grade, and regularly promoted students. Although retained and nonretained students were matched on gender, grade, age, self-concept, and level of motivation, the retainees were acknowledged to differ substantially from their promoted peers on unspecified sociological and psychological variables. Twenty-two students were retained in Grades one or two; 15 students were required to repeat either Grade three or four; 10 pupils in either Grades seven or eight were also retained. A comparable number of borderline or regularly promoted students were matched with the retained children in each of the three broad grade categories. Students were retained in the spring of 1983. All students whether retained or promoted were again tested in the spring of 1984. Retained pupils in the primary and intermediate elementary grades evidenced significant increases in reading, language, and mathematics while the promoted borderline students showed a decrease in achievement during the same year. Among the seventh and eighth Grade students, however, changes in achievement scores were indistinguishable between retained and the borderline placed pupils. Pomplun suggested that retention was more beneficial in the earlier elementary Grades than in middle schools. A limitation of the analyses is that students were followed for only one year.

Table 3. Characteristics of Additional Retention Studies

Author(s)	Quality of Controls	Outcome Pretest	Ability Pretest	Statistical Control for Outcome Pretest	Statistical Control for Ability Pretest	Type of Comparison	Grade Equivalent Units	Authors' Conclusion Effect of Retention
Cosden et al. (1995) (36-36)	Inadequate	No	No	No	No	Same-Grade	No	Negative
Pomplun (1988) (47-47)	Adequate	Yes	Yes	Yes	Yes	Same-Grade	No	Positive
Southard and May (1996) (66-66-66)	Adequate	Yes[a]	Yes[a]	Yes	Yes	Same-Grade	No	Negative[b]
Lorence et al. (2002) (736-28,351)	Adequate	Yes[a]	No	Yes	No	Same-Grade	No	Positive
Jacob and Lefgren (2004) (8,120-5,018)	Adequate	Yes[a]	No	Yes	No	Same-Age	No	Positive

[a] Retained and promoted students were significantly different on this variable at time of retention.
[b] Data indicate that retention had a positive effect on academic achievement.

Another study (Southard and May, 1996) incorporating measures of student ability prior to retention was based on students from three elementary schools in a suburban New York district. Kindergarten teachers rated 66 pupils (from 1982 to 1985) as being unable to perform first Grade work. These students were placed in pre-first grade transition classrooms. Several different comparison groups were analysed. One was a group of 24 students who entered kindergarten at the same time as the transitioned pupils. However, after being promoted to first Grade, these 24

children were required to repeat the grade. Another control group consisted of 66 regularly promoted students who had not been placed in a transition room nor retained in first Grade. Non-retained first Grade classmates of the 66 pre-first grade transitioned students composed the final control group. All students had taken the California Achievement Tests in listening and mathematics at the end of kindergarten. These variables were used as covariates to assess differences in group performance at the end of the first Grade year because both the transitioned and eventual first Grade retainees had significantly lower mean listening and mathematics scores than the regularly promoted pupils. Initial CAT scores were not the primary basis for retention; instead, student behaviour and classroom skills in kindergarten were the sources of placement recommendation. Achievement outcomes were taken from the reading and mathematics sections of the Iowa Tests of Basic Skills when all students were in first, second, fourth, and fifth Grade. Using the kindergarten listening scores as a covariate, the regularly promoted students and year-younger classmates of the transitioned students evidenced significantly higher reading scores than the children who had been placed in the pre-first grade transition room. Average reading and mathematics scores of the transitioned and eventual first Grade repeaters were comparable. However, children in the pre-first grade transition class reported significantly higher math scores than the regularly promoted and younger aged classmates at the end of first Grade. Nonetheless, the authors believed their findings implied that there was little long-term benefit to retention.

The authors acknowledged that "Advocates of pre-first programs might claim that these mixed results between the pre-first and comparison group students were evidence in support of these programs" (Southard and May, 1996, p.139). But the authors reached a negative conclusion because they assumed that the transitioned students should perform better than their younger classmates and the students who had to repeat first Grade. An alternative interpretation is that the transition year helped the academically challenged kindergartners catch up with the continually promoted students who demonstrated higher levels of mathematical ability. Had the authors been able to measure reading aptitude, instead of facility in listening at the end of kindergarten, the reading performance of the transitioned students may have been similar to that of the regularly promoted children in the later grades.

Another study reporting that grade retention is associated with a positive impact on school performance appeared a year after the publication of Jimerson's meta-analysis. Lorence, Dworkin, Toenjes, and Hill (2002) tracked the academic progress of a cohort of all low-performing third Graders in Texas public schools from 1994 through 1999. The same-grade scores of over 700 students who failed the state's mandatory reading test (the Texas Assessment of Academic Skills Test – TAAS), and had to repeat third Grade, were compared with the test results of over 27,000 socially promoted third Graders who also failed the TAAS reading examination. The retained third Graders not only caught up with the promoted students (who actually had higher reading scores than the retained at the end of third Grade), but the retainees statistically surpassed their low-performing promoted counterparts each year after being held back in third Grade. After statistically adjusting for initial performance levels and socio-economic variables, the retainees were outperforming the socially promoted pupils by 0.30 standard deviations at the end of seventh Grade. Although many of the socially promoted third Graders eventually passed the state reading test, the retained students, on average, passed the test a year earlier. The authors found no evidence that grade retention hindered the academic performance of students over the six years of study.

Two economists (Jacob and Lefgren, 2004) recently evaluated the impact of third Grade and sixth Grade retention among children enrolled in the Chicago Public School System during the mid and late 1990s. Using a regression-discontinuity design which statistically controlled for student prior test performance, in addition to social and demographic characteristics of the children and their neighbourhoods, the authors found that retention helped improve the performance of third Graders

in both reading and mathematics. However, retention was associated with somewhat lower scores in reading and mathematics among students who had been required to repeat sixth Grade. The authors caution that the estimated effect of grade retention was confounded with high stakes tests in certain grades which may partially account for the variation in scores by retention grade. Same-year tests were the basis of comparison for retained and promoted Chicago students, but minimum passing scores are required for exams in Grades three, six, and eight before being advancing to the next grade. For example, the retained sixth Graders took the seventh Grade ITBS test (a low-stakes test with no consequence for their promotion to eighth Grade) the same year when the promoted sixth Graders were required to take the eighth Grade ITBS test. The retained sixth Graders were probably less motivated to do well that year than the promoted sixth Graders who needed a minimum score on the eighth Grade exam before being promoted to Grade nine. In spite of the difficulty of interpreting the impact of grade retention, the authors did not find that making low performing students repeat a grade adversely affected their later academic achievement.

DISCUSSION

The Holmes (1989) meta-analysis is the source most often cited as demonstrating the futility of making students repeat a grade to improve their academic shortcomings. Since 1989 it has not been possible to read a discussion on grade retention which does not mention, in some form, Holmes summary that "The weight of evidence argues against grade retention" (p.28). Critics of grade retention practices (for example, Dawson, 1998) maintain that "the most valid and best designed studies…clearly supports promoting underachieving students over retaining them" and then cite the Holmes meta-analysis to support their assertion. Jimerson's (2001) meta-analysis will also be cited as a seminal study which discredits the practice of retention. Both Holmes and Jimerson are to be commended for their efforts to synthesize studies examining grade retention. However, the major conclusion of the present review is that findings pertaining to the effect of grade retention on student academic performance are not unequivocal; the issue of grade retention has not been resolved. Contrary to the conventional wisdom among educational researchers, this review of the extant grade retention literature argues ***there is no overwhelming body of scientifically sound evidence demonstrating that making academically challenged students repeat a grade is ineffective or harmful.*** This paper's conclusion derives from a detailed appraisal of specific studies listed in the two most recent comprehensive meta-analyses of grade retention. A meticulous examination of published studies indicates that the overall quality of research is characterized by serious methodological weaknesses. In-depth inspections of the Holmes and Jimerson meta-analyses reveal considerable shortcomings. Both of these summaries exhibit the many kinds of errors identified by Dunkin (1996) which commonly appear in meta-analyses. The vast majority of studies which conclude that retention is an ineffective educational practice contain so many limitations that inferences from them are highly questionable if not unwarranted.

Nonetheless, critics of retention argue that the quality of research is sufficiently high enough to discount the practice of making students repeat a grade. The most extreme example of this position is illustrated by Reynolds (1992), who maintains that the quality of findings appearing in the Holmes meta-analysis and more recent research is comparable to those from medical studies:

> In medicine, treatments that are shown to be ineffective or to have serious unintended effects do not gain approval from governmental bodies and are subsequently discarded or substantially revised to eliminate their undesirable effects. Despite the accumulated evidence to date, however, retention as an educational treatment has not followed such established scientific traditions. (p.118)

A glaring error in this assertion is the presumption that retention studies have attained the same high standards of research quality to which medical researcher must adhere. Unlike individuals

who conduct medical studies, educational researchers have not had the ability to randomly assign participants to specific narrow treatment conditions. Randomization is a powerful tool which can help rule out alternative explanations for causal effects in medical studies. To imply that retention studies use research designs comparable to the randomized experiments which dominate the field of medical research greatly exaggerates the extent to which results from retention research can be generalised. As Jackson (1975, p.624) noted in his overview of retention, the last time randomization was used to assign students to repeat a grade was over 60 years ago.

It should be evident from this review of both the Holmes (1989) and Jimerson (2001) meta-analyses that grade retention studies in no way approach the precision of medical research. It is inappropriate to aggregate findings from retention analyses based on inadequate controls and then assume the summary calculations will yield valid conclusions. Few studies examined from these meta-analyses meet acceptable criteria required for reasonable inferences. Valid inferences can not be made from the vast majority of studies in the aforementioned meta-analyses because the two groups of students did not have comparable characteristics when the decision to retain was made. Only four of the ten published articles in the Holmes summary would meet conventional criteria for achieving some degree of comparability between retained and promoted students; however, the quality of the two articles reaching negative conclusions about the usefulness of retention is suspect because of uncertainty in the nature of student differences at the time of retention. When compared to the Holmes review, Jimerson's meta-analysis fares somewhat better because 10 of the 18 studies appear to attempt to control for possible differences between retained and nonretained students, although the comparisons in 4 of these 10 studies are questionable. Rather than relying on a matching procedure to help equalize initial differences between retained and nonretained students, as did many of the articles cited in the Holmes summary, more of the authors in Jimerson's survey utilized regression procedures to make statistical adjustments. Southard and May (1996, p.141) specifically noted the importance of incorporating preretention indicators of outcome measures as covariates to obtain a more accurate assessment of making students repeat a grade; they declared their results would have been very different had they not statistically adjusted initial differences between retained and promoted students.

Even if some of the studies utilized better controls, the small sample sizes and the limited representativeness of the students can yield only the weakest of inferences about making students repeat a grade. Five of the 10 better designed studies analysed small numbers of observations and none of the 10 studies were based on national representative samples of the school population. Findings from the overwhelming majority of retention studies do not support the position that grade retention is an inappropriate remediation practice. This summary is consistent with the position of Alexander et al. (1994) who maintain that the strong opinions individuals hold regarding the impact of grade retention are actually based on weak empirical evidence from poorly designed studies.

In spite of these methodological shortcomings, opponents of making students repeat a grade will likely continue to cite the Holmes and Jimerson meta-analyses as authoritative proof that retention is an unsuitable educational practice. Nonetheless, the current review has identified seven studies which indicate that making students repeat a grade is associated with higher academic performance. The quality of the research designs of these positive studies are at least comparable to those alleged tightly controlled studies cited in Holmes (1989). Moreover, nine other studies in which the authors do not favour retention present findings which suggest that retention may result in some positive academic outcomes. Whether critics of grade retention will accept these positive studies as "demonstrating the effectiveness of retention as an intervention facilitating subsequent academic success" which Jimerson (1999, p.265) and other critics of retention (see also Holmes and Matthews, 1984, p.232) have required proponents of the practice to provide is unlikely.

Alexander et al. (2003, pp.16-20) contend that educational researchers have such strong negative opinions on the subject of grade retention that they are biased against any evidence which contradicts the view that holding students back a year in grade is a bad practice. Partial support for this view is evident in the kinds of criticisms made against studies which favour retention. Tanner and Galis (1997) have also suggested that "Even when the research is refereed and published we find that sometimes the reporting appears to be biased and misleading" (p.110). For example, Shepard, Smith, and Marion (1996) published a very detailed critique of the Alexander et al. (1994) study. Shepard and her colleagues speculated at great length on shortcomings and alternative interpretations of the Baltimore results. They concluded that the findings of Alexander et al. (1994) were flawed because of the manner in which scores were scaled; special education students should have been removed from the analyses; selection effects and regression artifacts resulted in erroneous findings.[15] Likewise, Shepard (2002) is highly critical of the Texas study (Lorence et al., 2002) which indicated that grade retention may have helped improve the academic competencies of low-performing elementary school children. Her major arguments are that: the authors did not adequately control for the special education students; improvement in test scores was largely the result of regression to the mean effects; selection effects yielded biased results; and public school teachers teach only information pertaining to the state's mandatory accountability test. My point is not that studies in favour of retention are beyond criticism; but that a more rigorous standard of what constitutes "acceptable research" is applied to studies which contradict the prevailing view on grade retention. Studies which conclude that retention is an ineffective remediation practice are not subjected to the higher standards of methodological rigor required for studies which are critical of letting academically challenged students automatically proceed to the next grade level.

The major purpose of this review, as is indicated in the opening paragraphs, is to draw attention to the situation that most of the studies reported in the research literature on retention are not sufficiently sound to support the claim that grade repetition is always wrong. More recent research suggests that grade retention may indeed help improve student learning. If thoughtful readers will make more of an effort to judge the evidence on retention objectively, this review will have served its purpose. Over a decade ago Kaestle (1993) discussed the "awful reputation of education research." He pointed out that policies and practices were sometimes highly politicised because the research did not allow the educational community to reach a consensus on important issues. The same can be said about educational researchers who disagree on the utility of making students repeat a grade. Although most educational researcher believe that making students repeat a grade is ineffective, this review challenges that position. A detailed examination of retention studies pertaining to academic achievement indicates there is no cumulative research that yields a firm conclusion on the topic. Instead of lauding research which supports a specific view and criticizing research which contradicts a favoured perspective, a better approach is to seek consensus as to what appropriate criteria are needed to determine if retention is worthwhile and then perform the necessary research which will provide answers. Karweit's (1992) earlier attempt to raise such issues has largely been ignored. For example, there seems to be little agreement as to what the goals of retention should be. Must retained students surpass the academic achievement of regularly promoted peers to be successful, or will merely catching up with socially promoted low-performing students met the criterion of effectiveness? Researchers have also differed in their definitions of retention and the kinds of educational practices which occur during the repeated year. These and other issues need to be addressed before judging the usefulness of retention practices.

[15] See Alexander et al. (2003, pp.265-279) for a detailed rebuttal to the criticisms of Shepard et al. (1996).

SUGGESTIONS FOR FUTURE RESEARCH

Several dimensions of retention research must be attended to if a more definitive assessment of the impact of making students repeat a grade is to emerge. One issue is that the impact of retention may vary by grade. To illustrate, Peterson et al. (1987) found that the learning gains obtained among second and third grade retainees persisted over time while the initial improvement observed among the retained first graders declined. Pierson and Connell (1992) reported that students retained from third to sixth grade experienced learning gains compared to matched counterparts. The Baltimore study found that children held back in first grade, or who were eventually placed in special education, did not appear to benefit as much from retention as those students retained in second or third grade. Alexander et al. (2003) suggest that children with the most noticeable learning disabilities are the first to be retained, usually in kindergarten or first grade. Such children are often much farther behind their peers in academic ability than children retained in later grades. Pupils retained in second or third grade were probably closer in academic ability to their promoted classmates. Because they did not have the kinds of problems typical of earlier retainees, the second and third grade repeaters were better able to learn the material during their year of retention. These findings imply that holding children in kindergarten or first grade may not lead to the same kinds of increased test scores observed among students in higher elementary grades. Pomplun's (1988) study of Florida students further reveals that retaining high school students will be less effective than requiring low achieving students in primary and intermediate grades. Poorly performing high school students may be so far behind their promoted classmates that making them repeat a grade will not enable them to complete a degree. Whether differences on the impact of retention across grades results from differences in the abilities of students held back in various grades should be examined in more detail.

A second issue requiring further research is disentangling the effect of retention with the specific instructional practices provided during the retention year. Karweit (1992) listed several kinds of educational practices offered during a repeated grade. Probably the most common is recycling students through the same grade with no additional resources or special assistance. The students simply repeat the same curriculum. This form of retention may not be helpful. A case in point is the Chicago Longitudinal Study Data, the basis of the three Reynolds' papers on grade retention which are often cited as showing that making students repeat a grade is ineffective. Thus far no one has suggested that the negative effects of retention observed among Chicago public school students may result from ineffective educational practices endemic to that specific school system. During the 1980s through the mid-1990s, Chicago was reputed to have one of the worst public school systems in the nation (see Hess, 1995; Vander Weele, 1994). Chicago school teachers may have been less concerned about helping remediate low-performing students than occurred in other school districts across the country. Indeed, Reynolds et al. (1997) describe the nature of retention practices in Chicago as follows: "Once students are retained, however, they usually get no special help with their schooling. They are often placed in low academic tracks only to repeat the previous year's instruction and ultimately disengage from school" (p.36). Merely repeating the failed grade may not help a student, but school districts more responsive to the needs of academically challenged students should be examined because the context of the school system may also affect student academic performance independent of making students repeat a grade.

Instead of merely recycling low achieving students through the same grade, another retention practice is to require students to repeat a grade, but also provide them additional learning opportunities during the year of retention. Several studies indicate that providing additional assistance and special programs to retained students may be more beneficial than only repeating the same curriculum. The longitudinal study of children in Mesa, Arizona (Peterson et al., 1987) showed that elementary school children whose teachers developed individual educational plans to address the retained students' academic shortcomings maintained higher scores than their socially

promoted counterparts. In their analyses of Minnesota children Jimerson et al. (1997) also "noted that many students in the retained group received additional academic services the repeated year" (p.21). The authors speculated that this extra assistance may have partially accounted for the higher math performance of the retained students. Likewise, Lorence et al. (2002, p.44) commented that interviews with Texas teachers and educational administrators revealed that Texas elementary students who were required to repeat a grade often received various forms of additional educational assistance. These examples imply that retention itself is not likely to be an effective remediation strategy when students are only recycled through the same educational program of the failed year. However, when combined with strategies focusing on the unique academic weaknesses of students, retention may help raise the achievement levels of low-performing pupils. These findings contradict the negative view that "the effects of most retention plus remediation approaches are likely to be disappointing" (McCoy and Reynolds, 1999, p.295).

A third kind of retention practice is placing academically challenged children in an alternative program, often a transition classroom, before actually classifying the students as failing. These transitional classes are usually created for kindergartners or children in first grade. However, the specific instructional practices available to children in these transitional classrooms are seldom described. One exception is Leinhardt's (1980) study of low-achieving kindergartners. Although this study has been cited as indicating that extra educational assistance combined with retention is ineffective, the special instruction offered the retained children placed in a transitional room was of much lower quality than provided socially promoted peers. Before making general conclusions about the effectiveness of retention combined with special help, more specific information is required to learn the specific retention policies students are subjected to and how they are implemented. Researchers should attempt to identify the kinds of specific educational activities occurring during the retention year, instead of grouping different kinds of instructional practices as referring to retention in only a generic sense.

Even if the specific educational practices occurring during retention can be identified, a major weakness of retention research is ascertaining the causal effect of making students repeat a grade. As previously mentioned, random assignment of students to different treatment conditions is considered the most powerful research design to assess causality. The implausibility of randomly making academically challenged students repeat a grade leaves researchers only quasi-experimental designs to control for differences between retained and promoted students. The two types of general procedures appearing in the retention literature are matching and statistical adjustments. Although findings are considered to be superior if retained and promoted students are well matched, or if important variables known to be related to academic success can be incorporated into a regression equation, both of these procedures may still result in biased outcomes. One shortcoming pertains to the issue of regression artifacts. The second is referred to as the problem of "omitted" or "unmeasured" variables associated with selection biases.

All educators acknowledge that students of similar ability may not obtain identical scores on an examination. Students have good days and bad days when taking tests. Students with very low scores one year will likely have higher scores on the same test the following year. Similarly, children with extremely high scores one year will probably have somewhat lower scores when the exam is taken again. This general phenomenon is commonly referred to as "regression to the mean." Campbell and Kenny (1999) caution that the process of matching or the use analysis of covariance methods may lead to findings that are simply regression artifacts. Studies of teacher initiated retention are problematic because teachers are more likely to make the lower performing students repeat a grade. A low performing retained pupil who is matched with a promoted student with a similar score may appear to be similar, but they still could have different levels of knowledge and ability. The retained pupil's poor test results could be due to an abnormally low test score resulting from transient idiosyncratic factors during the day of the examination. The test

results on which the teacher bases the retention decision may be due to a student having a low score which underestimates the child's true ability. Students retained under such circumstances will likely obtain appreciably higher scores on next year's exam because the initial test score was below the true level of student learning. Had the students required to repeat a grade been promoted, their test scores would still have been higher due to the regression to the mean phenomenon. Shepard (2002) and her colleagues (Shepard et al., 1996) argue that findings showing retention is associated with higher test scores result from regression artifacts. Critics of grade retention contend that the better test results observed among the retained students after being required to repeat a grade are mainly attributable to the regression to the mean, rather than better comprehension of the material. Statistical adjustments, such as entering the test score prior to the year of retention into a linear prediction model, may not adequately control for the regression effect. The lowest performing students, who are most likely to be retained, will probably report higher test scores at the end of the retention year.

Several methods have been suggested to determine the degree to which regression artifacts may be present in data. If test scores are available from several exams prior to retention, one can examine the test score means plotted over time to ascertain if student performance has been consistently declining. The presence of a steep decline in test results prior to the year of retention and a sharp increase following the grade repeated would imply a regression to the mean effect. If data prior to the year of retention exist, Campbell and Kenny (1999, pp.158-163) suggest performing a "time-reverse analysis" in which the values of the dependent variable and its covariate be reversed. That is, the value of the outcome measure after the retention year (T_2) would become the independent variable in a regression analysis while the initial value of the covariate (the value in T_1) becomes the dependent variable. Should the sign of the binary treatment variable reverse itself but remain similar in magnitude to that from the initial regression equation, a regression artifact is highly unlikely. The original estimated effect of the treatment is probably unbiased. Although critics of grade retention often argue that the positive effects of retention are due to regression artifacts, Campbell and Kenny (1999, pp.74-75) suggest that statistical adjustments likely underestimate the effect of treatments, especially when the pretest mean of the group given the treatment is smaller than that of the pretest value for the control group. Insofar as, prior to being retained, the mean test scores of students required to repeat a grade are almost always lower than those of the promoted students, analysis of covariance adjustments probably yield estimates lower than the true impact of the retention year on student academic performance.

An exacting methodologist could argue the major flaw of all retention research is that none have adequately addressed the issue of unmeasured or unspecified variables which affect the decision to retain a student. Some pupils are retained for academic reasons while others are held back because of a lack of emotional or behavioural maturity, which teachers or principals assume will retard future learning in the next grade. Those factors which lead to retention may also affect academic learning outcomes. These unspecified causes will often result in a teacher recommending one student repeat the grade while another student with a similar test score will be allowed to progress to the next grade. If these unmeasured variables affect both the retention decision and performance on tests in later grades, the calculated net effect of grade retention will be biased, often referred to as the problem of "omitted variables." Although matching or the use of test scores prior to retention as covariates will likely yield more accurate estimates of the impact of retention on academic achievement than not attempting to control for initial differences between retained and promoted children, neither procedure will result in unbiased findings.[16]

---

[16] Whereas the present paper concentrates exclusively on studies examining the impact of retention on academic achievement, another body of research focuses on the effect of grade retention and the completion of high school. The quality of research in these studies is generally higher than in the achievement literature. The consistency of findings

Over the last twenty years statisticians and econometricians have developed statistical procedures which can be used to help better address the problem of unmeasured variables. These techniques should allow educational researchers to obtain more accurate estimates of how making students repeat a grade, or any other educational intervention, influences their academic achievement. The current conceptual framework recommended to estimate causal effects is referred to as the "counterfactual account of causality." Although this approach is highly technical in nature, Winship and Morgan (1999) provide a general overview of the basic issues and various analytical strategies which can be used to obtain more accurate assessments of the impact of educational practices on student outcomes. Only the most basic features of this approach is presented here. The counterfactual approach tries to estimate the effect of being placed in one group as opposed to another. In natural experiments individuals are nonrandomly assigned to a treatment group or a control group. For example, students would have an observable outcome measure if retained or promoted. The counterfactual approach attempts to answer the question of what would happen to the children who were retained if they were instead promoted to the next grade. One could also ask what the consequences of repeating a grade would be for students who were promoted to the next grade. Although students have potential outcome in either state, an outcome can be measured in only one state. For example, a retained student has an observed score when retained and an unobserved counterfactual outcome if placed in the control or promoted group. In the current context, those factors which lead to a student repeating a grade or being promoted will likely be associated with a student's later academic achievement. Given that assignment to the treatment or control group will be correlated with the outcome variable of interest, standard ordinary least squares regression methods will not yield consistent estimates of the retention effect.

Numerous strategies have been suggested to reduce the correlation between the treatment and the outcome measure caused by assignment to the control or experimental group. An often used strategy is to create a "propensity score" or the probability that a person with certain characteristics will be assigned to the treatment group. One would use a large number of variables (Z_i) to compute the propensity score, $P(Z_i)$ of ending up in the treatment condition (here the retained group). The propensity score is like a mega-covariate for placement in the treatment group. Rosenbaum and Rubin (1985) then recommend matching the propensity score of a retained student with the closest propensity score of a promoted student in the control group. This approach assumes, however, that all variables which influenced placement into the treatment group are observed. If one assumes that unobserved or unspecified factors influence placement into the treatment or control group (a more reasonable assumption), Heckman (1978) suggests generating two variables, one for the likelihood of being placed in the treatment group and another for being selected into the control group. These two new selection effect variables can be entered into a regression equation along with the treatment variable and other control variables predicting the outcome measure of interest. Another strategy is to use instrumental variables which affect assignment into the experimental or control group, but does not directly influence the outcome measure. Although often used by economists, the instrumental variables approach suffers from certain shortcomings which may limit its usefulness (Winship and Morgan, 1999, pp.683-685).

---

from studies examining the impact of retention on school dropout behaviour suggests that making students repeat a grade will cause students to leave school without a diploma (Jimerson, Anderson, and Whipple, 2002). However, none of the studies examining the relationship between retention and dropping out of school have investigated the possibility that factors leading to retention also affect the likelihood of leaving school. Thus far, no one has investigated the possibility that models assessing the effect of grade retention on dropping out of school are misspecified because of unobserved variables affecting both retention status and school exit behaviour.

More rigorous analytical procedures to gauge the impact of retention on student performance require several measures of the outcome variable both prior to and following retention for students in the treatment and control groups are described by Winship and Morgan. The more advanced statistical modelling procedures suggested by the counterfactual approach to causality may yield new insights about the degree to which grade retention influences student learning outcomes.

Beginning with then President Clinton's (1998) appeal to end social promotion and culminating with the No Child Left Behind Act (2002) legislation, many state educational agencies and some local districts have mandated that students should be promoted only after mastering the basic skill requirements for their grade level. Political pressure to change school promotion polices will likely increase the number of children required to repeat a grade. Although critics have railed against the practice of retention, the present political climate in the United Sates will likely not change sufficiently in the near future to allow educational practitioners to modify recently implemented stringent promotion policies. However, the present situation may provide researchers an opportunity to more thoroughly evaluate retention practices. Whereas many previous studies were limited to a small number of students, the greater availability of retained students from more diverse racial and economic backgrounds should enable a better description of the consequences of retention. Given that not all districts will utilize the same remediation practices, the great variation in specific retention policies will allow researchers a better opportunity to identify which programs and strategies are more likely to enhance the academic standing of retained children. While such a suggestion may seem self-serving, educational researchers should take advantage of those situations which can help increase our understanding of retention practices and their implications for students. Even though many readers already have strong opinions on the retention issue, those individuals who acknowledge the limitations of the existing research should search for research settings which will enable us to further our comprehension of retention processes and their outcomes.

REFERENCES

Note: References marked with a single asterisk (*) indicate studies appearing in the Holmes (1989) meta-analysis. References with two asterisks (**) are cited in the Jimerson (2001) meta-analysis.

*Abidin, R. P., Golladay, W. M., and Howerton, A. L. (1971) Elementary school retention: An unjustifiable, discriminatory, and noxious educational policy. *Journal of School Psychology*, 9(4), 410-417.

**Alexander, K. L., Entwisle, D. R., and Dauber, S. L. (1994) *On the success of failure*: *A reassessment of the effects of retention in the primary grades*. New York: Cambridge University Press.

Alexander, K. L., Entwisle, D. R., and Dauber, S. L. (2003) *On the success of failure*: *A reassessment of the effects of retention in the primary grades* (2nd ed.). New York: Cambridge University Press.

Anfinson, R. D. (1941) School progress and pupil adjustments. *The Elementary School* Journal, 41(7), 507-514.

Archer, G. A. (1967) *A study of nonpromotion relative to normal grade expectancy in selected Catholic elementary schools in Illinois*. Unpublished doctoral dissertation, Loyola University, Chicago. Cited in C. T Holmes (1989). Grade level retention effects: A meta-analysis of research studies. In L. A. Shepard and M. L. Smith (eds) *Flunking grades: Research and policies on retention*. London: The Falmer Press.

Campbell, D. T. and Kenny, D. A. (1999) *A primer on regression artifacts*. New York: Guilford Press.

Campbell, D. T. and Stanley, J. C. (1966) *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

*Chansky, N. M. (1964) Progress of promoted and repeating grade 1 failures. *Journal of Experimental Education*, 32(3), 225-237.

Clinton, W. (1998) State of the Union Address (January 1, 1999). Washington, DC: U.S. Government Printing Office.

Cohen, J. (1977) *Statistical power analysis for the behavioral sciences* (Revised ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Coleman, J. S. and Karweit, N. L. (1972) *Information systems and performance measures in schools.* Englewood Cliffs, NJ: Educational Technology Publications. [The Rand Corporation]

Cosden, M., Zimmer, J., Reyes, C., and del Rosario Gutierrez, M. (1995) Kindergarten practices and first-grade achievement for Latino Spanish-speaking, Latino English-speaking, and Anglo students. *Journal of School Psychology,* 33(2), 123-141.

Darling-Hammond, L. and Falk, B. (1997) Using standards and assessments to support student learning. *Phi Delta Kappan*, 79(3), 190-199.

Dawson, P. (1998). A primer on student grade retention: What the research says. *NASP Communique*, 26(8), 28-30. Retrieved June 18, 2004, http:www.nasponline.org/publications/ cq268retain.html.

**Dennebaum, J. M. and Kulberg, J. M. (1994) Kindergarten retention and transition classrooms: Their relationship to achievement. *Psychology in the Schools*, 31(1)*,* 5-12.

*Dobbs, V. and Neville, D. (1967) The effect of nonpromotion on the achievement of groups matched from retained first graders and promoted second graders. *Journal of Educational Research*, 60(10), 470-475.

Dunkin, M. J. (1996) Types of errors in synthesizing research in education. *Review of Educational Research*, 66(2), 87-97.

**Ferguson, P. C. (1991) Longitudinal outcome differences among promoted and transitional at-risk kindergarten students. *Psychology in the Schools*, 28(2), 139-146.

**Ferguson, P. C. and Mueller-Streib, M. (1996) Longitudinal outcome effects of non-at-risk and at-risk transitions first-grade samples: A follow-up study and further analysis. *Psychology in the Schools*, 33(1), 38-45.

**Hagborg, W. J., Masella, G., Palladino, P. and Shepardson, J. (1991) A follow-up study of high school students with a history of grade retention. *Psychology in the Schools*, 28(4), 310-317.

Harrington-Lueker, D. (1998) Retention vs. social promotion. *The School Administrator*, 55(7), 6-12.

Heckman, J. J. (1978) Dummy endogenous variables in a simultaneous equation system. *Econometrica*, 46(4), 931-61.

Hess, A. G. (1995) *Restructuring urban schools*. New York: Teachers College Press.

Heubert, J. P. and Hauser, R. M. (1999) *High stakes: Testing for tracking, promotion, and graduation*. Washington D.C.: National Academy Press.

Holmes, C. T. (1989) Grade level retention effects: A meta-analysis of research studies. In L. A. Shepard and M. L. Smith (eds) *Flunking grades: Research and policies on retention*. London: The Falmer Press.

Holmes, C. T. and Matthews, K. M. (1984) The effects of non-promotion on elementary and junior high school pupils: A meta-analysis. *Review of Educational Research*, 54(2)*,* 225-236.

Jackson, G. B. (1975) Research evidence on the effects of grade retention*. Review of Educational Research*, 45(4), 613-635.

Jacobs, B. A. and Lefgren, L. (2004) Remedial education and student achievement: A regression-discontinuity analysis. *The Review of Economics and Statistics*, 86(1), 226-244.

Jimerson, S. (2001) Meta-analysis of grade retention research: Implications for practice in the 21st century. *School Psychology Review*, 30(3), 420-437.

**Jimerson, S. (1999) On the failure of failure: Examining the association between early grade retention and education and employment outcomes during late adolescence. *Journal of School Psychology*, 37(3), 243-272.

Jimerson, S. (2004) Is grade retention educational malpractice? Empirical evidence from meta-analyses examining the efficacy of grade retention. In H. Walberg, A. J. Reynolds and M. C. Wang (eds) *Can Unlike Students Learn Together?* Greenwich, CT: Information Age Publishing, Inc.

**Jimerson, S., Carlson, E., Rotert, M., Egeland, B., and Sroufe, L. A. (1997) A prospective, longitudinal study of correlates and consequences of early grade retention. *Journal of School Psychology*, 35(1), 3-25.

Jimerson, S., Anderson, G. E., and Whipple, A. D. (2002) Winning the battle and losing the war: Examining the relationship between grade retention and dropping out of high school. *Psychology in the Schools*, 39(4), 441-457.

**Johnson, E. R., Merrell, K. W., and Stover, L. (1990) The effects of early grade retention on the academic achievement of fourth-grade students. *Psychology in the Schools*, 27(4), 333-338.

*Kamii, C. K. and Weikart, D. P. (1963) Marks, achievement, and intelligence of seventh graders who were retained (nonpromoted) once in elementary school. *Journal of Educational Research*, 56(9), 452-459.

Kaplan, R. M. and Saccuzzo, D. P. (1997) *Psychological Testing*: *Principles, applications, and issues* (4th ed.). Pacific Grove, CA: Brooks/Cole Publishing Company.

Kaestle, C. F. (1993) The awful reputation of educational research. *Educational Researcher*, 22(1), 25-31.

Karweit, N. L. (1992) Retention policy. In M. C. Alkin (ed) *Encyclopedia of Educational Research* (6th ed.), New York: Macmillan.

Karweit, N. L. (1999) *Grade retention: Prevalence, timing, and effects* (Report No. 33). Baltimore: Johns Hopkins University, Center for Research on the Education of Students Placed at Risk.

*Leinhardt, G. (1980) Transition rooms: Promoting maturation or reading education? *Journal of Educational Psychology*, 72(1), 55-61.

Lorence, J., Dworkin, A. G., Toenjes, L. A., and Hill, A. N. (2002). Grade retention and social promotion in Texas, 1994-1999: Academic achievement among elementary school students. In D. Ravitch (ed) *Brookings Papers on Education Policy 2002*. Washington DC: The Brookings Institution.

Mabry, L. (1995) Review of the Metropolitan Readiness Tests, 5th ed. In J. C. Impara and J. C. Conoley (eds) *Twelfth mental measurements yearbook*. Lincoln, NE: The Buros Institute of Mental Measurements, University of Nebraska-Lincoln.

**Mantzicopoulos, P. Y. (1997) Do certain groups of children profit from early retention? A follow-up study of kindergartners with attention problems. *Psychology in the Schools*, 34(2), 115-127.

**Mantzicopoulos, P. and Morrison, D. (1992) Kindergarten retention: academic and behavioral outcomes through the end of second grade. *American Educational Research Journal*, 29(1), 182-198.

*May, D. C. and Welch, E. L. (1984) The effects of developmental placement and early retention on children's later scores on standardised tests. *Psychology in the Schools*, 21(3), 381-385.

Maxwell, S. E. and Delaney, H. D. (1990) *Designing experiments and analyzing data*. Belmont, CA: Wadsworth Publishing.

Maxwell, S. E., O'Callaghan, M. F., and Delaney, H. D. (1993) Analysis of covariance. In L. K. Edwards (ed) *Applied analysis of variance in behavioral science*, New York: Marcel Dekker, Inc.

**McCombs Thomas, A., Armistead, L., Kempton, T., Lynch, S., Forehand, R., Nousiainen, S., Neighbors, B., and Tannenbaum, L. (1992) Early retention: Are there long-term beneficial effects? *Psychology in the Schools*, 29(4), 342-347.

**McCoy, A. R. and Reynolds, A. J. (1999) Grade retention and school performance: An extended investigation. *Journal of School Psychology*, 37(3), 273-298.

**Meisels, S. J. and Liaw, F. (1993) Failure in grade: Do retained children catch up? *Journal of Educational Research*, 87(2), 69-77.

National Association of School Psychologists [NASP]. (2003). Position statement on student grade retention and social promotion (revised April 12, 2003). Retrieved June 18, 2004, from http://www.nasponline.org/information/pospaper_graderetent.html.

*Niklason, L. B. (1984) Nonpromotion: A pseudoscientific solution. *Psychology in the Schools*, 21(4), 485-499.

No Child Left Behind Act of 2001, 20 U.S.C. § 6301 (2002).

Owings, W. A. and Magliaro, S. (1998) Grade retention: A history of failure. *Educational Leadership*, 56(1), 86-88.

*Peterson, S. E., DeGracie, J. S., and Ayabe, C. R. (1987) A longitudinal study of the effects of retention/promotion on academic achievement. *American Educational Research Journal*, 24(1), 107-118.

**Phelps, L., Dowdell, N., Rizzo, F. G., Ehrlich, P., and Wilczenski, F. (1992) Five to ten years after placement: The long-term efficacy of retention and pre-grade transition. *Journal of Psychoeducational Assessment*, 10(3), 116-123.

**Pierson, L. and Connell, J. P. (1992) Effect of grade retention on self-system processes, social engagement, and academic performance. *Journal of Educational Psychology*, 84(3), 300-307.

Pomplun, M. (1988) Retention, the earlier the better? *Journal of Educational Research*, 81(5), 281-287.

Potter, L. (1996). Examining the negative effects of retention in our schools. *Education*, 117(2), 268-270, 250.

**Reynolds, A. J. (1992) Grade retention and school adjustment: An explanatory analysis. *Educational Evaluation and Policy Analysis*, 14(2), 101-121.

**Reynolds, A. J. and Bezruczko, N. (1993) School adjustment of children at risk through fourth grade. *Merrill-Palmer Quarterly*, 39(4), 457-480.

Reynolds, A. J., Temple, J., and McCoy, A. (1997) Grade retention doesn't work. *Education Week,* 17(3) [September 17], 36.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, 39(1), 33-38.

**Rust, J. O., and Wallace, K. A. (1993) Effects of grade level retention for four years. *Journal of Instructional Psychology*, 20(2), 162-166.

*Sandoval, J. and Fitzgerald, P. (1985) A high school follow-up of children who were nonpromoted or attended a junior first grade. *Psychology in the Schools*, 22(2), 164-170.

Schuyler, N. B. and Matter, M. K. (1983) To retain or not to retain: Should achievement be your guide? Paper presented at the annual meeting of the American Educational Research Association, Montreal. Cited in C. T. Holmes (1989). Grade level retention effects: A meta-analysis of research studies. In L. A. Shepard and M. L. Smith (eds) *Flunking grades: Research and policies on retention*. London: The Falmer Press.

Seltzer, M. H., Frank, K. A., and Bryk, A. S. (1994) The metric matters: The sensitivity of conclusions about growth in student achievement to choice of metric. *Educational Evaluation and Policy Analysis*, 16(1), 41-49.

Shepard, L. (2000). Cited in Ending social promotion by Debra Viadero. *Education Week on the Web*. (March 15, 2000). Retrieved on June 1, 2004, from [Online] http://www.edweek.org/ew/ewstory.cfm?slug=27.social.h19andkeywords=Ending per cent20Social per cent20Promotion.

Shepard, L. (2002) Comment on grade retention and social promotion in Texas, 1994-1999: Academic achievement among elementary school students. In D. Ravitch (ed) *Brookings Papers on Education Policy 2002*. Washington DC: The Brookings Institution.

Shepard , L. (2004) Understanding research on the consequences of retention. In H. Walberg, A. J. Reynolds and M. C. Wang (eds) *Can Unlike Students Learn Together?* Greenwich, CT: Information Age Publishing, Inc.

*Shepard, L. and Smith, M. L. (1987) Effects of kindergarten retention at the end of first grade. *Psychology in the Schools*, 24(4), 346-357.

Shepard, L. A. and Smith, M. L. (1990) Synthesis of research on grade retention. *Educational Leadership*, 47(8), 84-88.

Shepard, L. A., Smith, M. L., and Marion, S. F. (1996) Failed evidence on grade retention. [Review of the book *On the success of failure*: *A reassessment of the effects of retention in the primary grades*.] *Psychology in the Schools*, 33(3), 251-261.

Southard, N. A. and May, D. C. (1996) The effects of pre-first grade programs on student reading and mathematics achievement. *Psychology in the Schools*, 33(2), 132-142.

Stoner, G. (1995) Review of the Metropolitan Readiness Tests, 5th ed. In J. C. Impara and J. C. Conoley (eds), *Twelfth mental measurements yearbook.* Lincoln, NE: The Buros Institute of Mental Measurements, University of Nebraska-Lincoln.

Tanner, C. K. and Galis, S. A. (1997) Student retention: Why is there a gap between the majority of research findings and school practice? *Psychology in the Schools*, 34(2), 107-114.

Thorndike, R., M. (1997) *Measurement and evaluation in psychology and education* (6th ed.). Upper Saddle River, NJ: Prentice Hall

U.S. Department of Education. (1999) *Taking responsibility for ending social promotion*. Washington, DC: U.S. Government Printing Office.

Vander Weele, M. (1994) *Reclaiming our schools: The struggle for Chicago school reform*. Chicago: Loyola University Press.

Wilson, M. (1990) [Review of the book *Flunking grades: Research and policies on retention*.] *Educational Evaluation and Policy Analysis*, 12(2), 228-230.

Winship, C. and Morgan, S. L. (1999) The estimation of causal effects from observational data. In K.S. Cook and John Hagan (eds) *Annual Review of Sociology Vol. 25*. Palo Alto, CA: Annual Reviews.

Wolf, F. M. (1986) *Meta-analysis: Quantitative methods for research synthesis*. Beverly Hills, CA: Sage.

Wright, J. B. (1979) The measured academic achievement of two groups of first grade students matched along five variables when one group has been retained. Unpublished doctoral dissertation, Loyola University, Chicago. Cited in C. T Holmes (1989). Grade level retention effects: A meta-analysis of research studies. In L. A. Shepard and M. L. Smith (eds) *Flunking Grades: Research and Policies on Retention*. London: The Falmer Press.