An Investigation of the Relationship Between Retention in First Grade and Performance on High Stakes Tests in Third Grade

Jan N. Hughes
Texas A&M University

Qi Chen
University of North Texas

Felix Thoemmes
Oi-man Kwok
Texas A&M University

The association between grade retention in first grade and passing the third grade state accountability tests, the Texas Assessment of Knowledge and Skills (TAKS) reading and math, was investigated in a sample of 769 students who were recruited into the study when they were in first grade. Of these 769 students, 165 were retained in first grade and 604 were promoted. Using propensity matching, we created five imputed datasets (average N = 321) in which promoted and retained students were matched on 67 comprehensive covariates. Using GEE models, we obtained the association between retention and passing the third grade TAKS reading and math tests. The positive association between retention and math scores was significant, whereas the association was marginally significant for reading scores.

Keywords: grade retention, high stakes testing, elementary students

Grade retention in U.S. schools has a long history characterized by fluctuations in the frequency and application of this educational intervention (Bali, Anagnostopoulos, & Roberts, 2005; Lorence, 2006; Owings & Magliaro, 1998). These fluctuations have been associated with, and presumably reflect, shifts in educators’ and policymakers’ beliefs about the effectiveness of grade retention and the conditions under which it should be applied. Because no institution or agency tracks national data on the frequency of grade retention, precise estimates of changes in frequency across decades are not available. According to the U.S. National Center for Education Statistics (2006), in 2004, 9.6% of youth ages 16 to 19 had ever been retained in grade. This represents a decrease from 16.1% in 1995. However, given wide variations across

This research was supported in part by a grant to Jan N. Hughes from the National Institute of Child Health and Development (5 R01 HD39367-02). Correspondence should be addressed to Jan N. Hughes, Department of Educational Psychology, Texas A&M University, MS4225, College Station, TX 77843-4225; e-mail: jhughes@tamu.edu.
states in both the overall frequency of grade retention and the policies that affect its use, these national trend data tell us little about the factors that affect use of this educational intervention.

Grade Retention in the Context of Accountability Testing

Considerable interest has focused on the effect of the standards-based reform movement on grade retention practices. The standards-based reform movement emphasizes setting competency standards for students at each grade level and holding both schools and students accountable for meeting them. The reform movement calls for an end to social promotion, the practice of allowing students who have failed to meet standards to advance to the next grade with their peers instead of completing or satisfying the requirement (U.S. Department of Education, 1999). The No Child Left Behind federal legislation passed in 2001 requires that assessments, aligned with state standards, be used to measure the achievement of all children at each grade level (U.S. Department of Education, 2006). The use of such tests for purposes of making decisions concerning promotion and graduation is referred to as high stakes testing (Heubert & Hauser, 1999).

States and school districts that have implemented policies requiring students to demonstrate mastery of grade-level academic competencies to advance to the next grade often report an increase in the rate of grade retention in the grades affected by the policy (e.g., Florida Department of Education, 2005; Texas Education Agency, 2007). Florida and Texas offer examples of implementation of policies requiring students to pass high stakes tests beginning in third grade in order to advance to the next grade. Implementation beginning in third grade is based on the reasoning that students in grades K–3 are learning to read and that beyond grade 4, students are reading to learn and that tests are less appropriate at younger ages (Heubert & Hauser, 1999). “Third-grade retention is centered on students’ ability to read proficiently and is a necessary intervention step to ensure students will be able to meet the more rigorous standards of subsequent grades” (Florida Department of Education, 2005, p. 1). Each state has also implemented policies requiring students to demonstrate grade-level standards in additional academic areas such as mathematics and science at higher grade levels.

Since 1999, Texas statutes have required schools to assess literacy of students in grades kindergarten through third grade using a state-approved measure and to provide remedial instruction to students who fail to demonstrate grade-level literacy competencies. Since the 2002–2003 year, students in grade 3 have been required to pass the state reading test to advance to grade 4. The test, known as the Texas Assessment of Knowledge and Skills (TAKS), is aligned with grade-level curriculum and assessment standards. Students are given three opportunities to pass the tests, and school districts are required to provide accelerated instruction in the subject areas failed after each test administration. Beginning in the 2004–2005 year, fifth grade students were required to meet performance standards for both reading and math in order to advance (Texas Education Agency, 2004). In the 1st year of implementation of the third grade promotional gate, there was a 14% increase in the percentage of students retained in grade (from 2.4% in 2001–2002 to 2.8% in 2002–2003). In the 1st year of implementation of the fifth grade promotional gate, the percentage of retained students increased more than 300% over the prior year, from 1% in 2003–2004 to 3.5% in 2004–2005 (Texas Education Agency, 2007).

In 2002–2003, Florida implemented a similar requirement that third graders demonstrate grade-level reading standards by the end of the school year. In the 1st year of this policy, 18.5% of students were retained in grade 3, an increase of 331% from the previous year (Florida Department of Education, 2005). In Florida and Texas, a substantial number of students who failed the third grade reading test have been promoted based on one or more exceptions (Florida Department of Education, 2004; Texas Education Agency, 2007). For example, of the grade 3 students in Texas who failed the TAKS or a state-required alternate reading assessment, only 44.4% were retained in grade 3 the next year. Of grade 3 students who passed the TAKS or the alternate reading assessment, 99.7% were promoted to grade 4 (Texas Education Agency, 2007).
Hughes et al.

It is important to recognize that the implementation of policies requiring passing performance on curriculum-aligned tests for promotion to the next grade is a part of comprehensive accountability policies that also includes consequences for schools and teachers based on student performance on curriculum-aligned tests. High stakes testing has resulted in closer monitoring of student competencies beginning in kindergarten and intense pressure on schools to report a high “passing rate” on the state accountability tests (Booher-Jennings, 2005; Watanabe, 2007). This pressure could result in an increase in retention in the early grades if educators reason that low achieving primary grade students will be more successful on the forthcoming state accountability test if they are given an extra year of instruction. This view would be consistent with educators’ belief that grade retention in the primary grade provides low achieving students an opportunity to “catch up” with their (younger) grade peers—a one-time adjustment that puts the students on a more favorable academic trajectory (Tomchin & Impara, 1992).

Data on the performance of early grade-retained students on subsequent state accountability tests are sparse. In the 2005–2006 year in Florida, of the 12,685 students taking the fifth grade FCAT Reading test who had failed the third grade FCAT Reading in 2002–2003 and repeated third grade, 60% obtained a passing score (i.e., level 2 or higher), compared with 84% of all fifth graders taking the FCAT Reading test in 2006 (Powell, 2008). Furthermore, approximately one third of students retained in third grade demonstrated reading proficiency above the lowest passing level (level 2, defined as limited success with curriculum) at fifth grade, less than half the rate when compared to all fifth grade students taking the fifth grade FCAT Reading in 2006 (Powell, 2008). Furthermore, approximately one third of students retained in third grade demonstrated reading proficiency above the lowest passing level (level 2, defined as limited success with curriculum) at fifth grade, less than half the rate when compared to all fifth grade students taking the fifth grade FCAT Reading in 2006. These data suggest that the majority of third grade retainees continued to perform below their grade-level peers 2 years post-retention. However, without a comparison group of students who were matched to the retainees on relevant variables (e.g., cognitive ability, behavioral adjustment, home learning environment) but who were promoted, these data cannot tell us the percentage of third grade retainees who would have passed the FCAT Reading, had they been promoted. To answer the question of whether grade retention is associated with performance on subsequent state standards-based tests, it is important to employ research designs that control for pre-retention differences on relevant variables.

Previous Research on Grade Retention Effects on Achievement

Researchers have attempted to assess the effects of grade retention on achievement for more than three decades (for meta-analytic reviews, see Holmes, 1989; Jimerson, 2001b; for narrative reviews, see Jimerson, 2001a; Shepard, Smith, & Marion, 1996; Sipple, Killeen, & Monk, 2004). The unanimous conclusion from these reviews is that grade retention offers few if any benefits to the retained student. For example, in a meta-analysis of 18 studies published from 1990 to 1999, Jimerson reported that the effect size of retention on achievement was –.39. However, most of the studies included in these reviews are plagued by significant methodological limitations, the most important being lack of a comparison group of promoted peers equivalent prior to retention on achievement and other variables predictive of achievement.

Recently, researchers have challenged the view that clear conclusions about the effect of grade retention are warranted, based on methodological limitations of extant studies (Hong & Raudenbush, 2005; Wu, West, & Hughes, 2008a). For example, Lorence (2006) criticizes prior meta-analytic studies by Holmes (1989) and Jimerson (2001b) for using a “score card” approach to counting the frequency of negative, positive, and nonsignificant effects or calculating weighted effect sizes without regard for the methodological quality of studies included in the meta-analysis. Lorence judged that only four studies reviewed by Jimerson used both adequate comparison groups and statistical controls.

Using multi-level modeling, Allen, Chen, Willson, and Hughes (2009) conducted a meta-analysis of 207 achievement effects nested in 22 studies published from 1990 to 2007 that met minimal criteria for control for selection effects. Study-level variables included the quality of the control for selection effects (i.e., control for pre-retention differences between promoted and retained students). As suggested by Lorence
Retention and Performance on High Stakes Tests

(2006), quality of control was associated with less negative (or, in some cases, more positive) effects. Specifically, studies employing adequate to good methodological designs yielded effect sizes not statistically significantly different from zero. Allen et al. (2009) also found that effect sizes differed based on whether retained and promoted students were compared when they were the same age or in the same grade; retention effects were less negative when same-grade comparisons were employed. Retained students often show a sharp improvement, relative to promoted peers, in meeting grade-level standards during the repeat year, when retained students are exposed to familiar curriculum; however, this improvement often disappears 2 to 3 years subsequent to retention (Alexander, Entwisle, & Dauber, 2003; Wu, West, & Hughes, 2008b). Some researchers have argued that same-grade comparisons are more consistent with the purpose of retention, which is to provide students the opportunity to be more successful in meeting the academic demands of future grades (Karweit, 1999; Lorence, 2006).

Previous studies investigating the effect of retention on achievement have employed a number of different measures of achievement, including grades, teacher ratings, and standardized tests. Scores on nationally standardized tests of achievement are the most commonly employed achievement outcome in studies included in published meta-analyses (Jimerson, 2001a). Grades and teacher ratings of performance not only introduce the possibility of teacher bias but also may be influenced by the class ability composition. However, if a nationally standardized test is not aligned with the school’s curriculum, it may not offer a valid index of the students’ success in meeting the academic challenges of their classrooms.

Because the standards-based movement aims to objectify and standardize decisions concerning promotion to ensure that students have the skills necessary to succeed at higher grades, tests such as the TAKS and FCAT that are aligned with the grade-level curriculum may offer a more direct answer to the question of whether grade retention achieves its purpose. However, to assess the effect of grade retention on subsequent performance on such tests, a study must use strong controls for pre-retention differences between promoted and retained students. As noted by Alexander et al. (2003), “This is a large order because promoted children who test at the same level as retained children are not likely to be in other ways the same: if they were the same in all relevant respects, they presumably would also be retained” (p. 30).

Propensity Scores

The most challenging design issue in studies of the effect of grade retention is making causal inferences about grade retention in the absence of a randomized experimental design (Campbell & Stanley, 1963; Shadish, Cook, & Campbell, 2002). Because students are not randomly assigned to the intervention (i.e., retention or promotion), a failure to adequately control for pre-existing differences between retained and promoted students that may affect students’ academic and social trajectories leaves open the possibility that pre-existing vulnerabilities rather than retention may be the cause of post-retention outcomes. Traditional methods of adjustment (e.g., entering a number of separate covariates as a block in regression analyses) are limited because they can use only a limited number of observed covariates (Shadish et al., 2002). To solve this problem, recently researchers have employed propensity scores to minimize the effect of pre-existing differences between retained and promoted students (Hong & Yu, 2008; Wu et al., 2008a). A propensity score is defined as the conditional probability of assignment to a particular treatment given a vector of observed covariates (Rosenbaum & Rubin, 1983). It is a scalar function of observed covariates that summarizes information required to balance the distribution of the covariates. Propensity score matching is a parsimonious way of reducing bias because it generates a single index—the propensity score—that summarizes information across potential confounds. Propensity scores can be used to reduce bias due to selection effects through matching, stratification, regression adjustment, or some combination of the three (Rosenbaum, 2002).

Study Purpose and Approach

The purpose of this study was to determine the association between repeating first grade and
passing the third grade TAKS Reading and Math tests for students attending one of three school districts in Texas (1 urban and 2 small city). Promoted students were in third grade in the 2003–2004 school year. Although passing the third grade Math test is not required for advancement to the next grade, passing rates on both Reading and Math tests at each grade level beginning at grade 3 are widely publicized and play the major role in determining the performance ratings for individual schools and for school districts (Texas Education Agency, 2008b).

As described in the Method section, we employ multi-level logistic regression to model school-level effects and propensity scores to control for differences between promoted and retained children on 67 variables assessed when all students were in first grade, prior to any student being retained. These variables were intended to be as comprehensive as possible, including variables that have been shown in prior research to be related to grade retention or to achievement in the elementary grades.

Method

Participants

Our strategy was to select an initial sample of children at high risk for retention and then to identify subsamples of retained and nonretained children within the larger sample for the generalized estimating equations (GEE) analysis that were closely matched on a comprehensive set of baseline variables. The initial sample of participants who are participating in our longitudinal study on the effect of grade retention was recruited from three school districts in Texas (1 urban and 2 small city) across two sequential cohorts in first grade during fall 2001 and 2002. School districts were chosen on the basis of their representation of the ethnic breakdown of urban areas of Texas. The student enrollment for the three schools was 42% White, 25% African American, 27% Hispanic, and 6% other. Children were eligible to participate in the longitudinal study if they scored below the median score on a state approved district-administered measure of literacy, spoke either English or Spanish, were not receiving special education services, and had not previously been retained in first grade. School records identified 1,374 children as eligible to participate. Because teachers distributed consent forms to parents via children’s weekly folders, the exact number of parents who received the consent forms cannot be determined. Incentives in the form of small gifts to children and the opportunity to win a larger prize in a lottery were instrumental in obtaining 1,200 returned consent forms, of which 784 parents (65%) provided consent and 416 declined.

Analyses of a broad array of archival variables available on all eligible children including performance on the district-administered test of literacy (standardized within district, due to differences in test used), age, gender, ethnicity, eligibility for free or reduced price lunch, bilingual class placement, cohort, and school context variables (i.e., % ethnic/racial minority; % economically disadvantaged) did not indicate any differences between children with and without consent. The resulting sample of 784 participants (52.6% male) closely resembles the population from which they were drawn on demographic and literacy variables relevant to students’ educational performance. The ethnic composition of the achieved sample (n = 784) was 37% Hispanic (39% of whom were Spanish language dominant), 34% Caucasian, 23% African American, and 6% other; 62% qualified for free or reduced cost lunch. The mean full scale IQ based on the Universal Nonverbal Intelligence Test (Bracken & McCallum, 1998) for the sample was 92.91 (SD = 18.01), and the mean reading achievement score was 96.40 (SD = 14.28).

Children were classified as retained in first grade if they were enrolled in a first grade classroom at the beginning of the following academic year (Year 2). Retention status was obtained from school district records and was available for 769 of the original 784 children. Of these children, 165 (21%) were retained in first grade and 604 were promoted to second grade. Propensity scores, the probability of being retained, were calculated for these 769 students (see below).

Assessment

Variables used to calculate propensity scores. The 67 variables used to compute propensity scores (see below) were collected when all (769) students were in first grade, prior to any child
being retained in grade. A complete list of the 67 baseline variables used in calculation of propensity scores is included in the appendix. These 67 variables were selected to be as comprehensive as possible, including variables that have been shown in prior research to be related to early grade retention or to early academic achievement. These variables came from teacher questionnaires, parent questionnaires, child interviews and child testing, school records, and peer sociometric testing. A summary of assessment procedures for these 67 variables follows. Details on assessment procedures are reported in Willson and Hughes (2009).

Demographic information including child age, ethnicity, economic adversity, bilingual placement, and status as limited English proficient (LEP) were obtained from school district records. Teachers and parents completed questionnaires that were mailed to them that assessed their perceptions of the child’s academic, behavioral, and social development. Parents also provided information on sociodemographic variables. Teachers and parents received $25.00 for completing and returning the questionnaires. Sociometric data were obtained via individual interviews with all students in classrooms (not just study participants) for whom parent permission was obtained.

In individual sessions, children were administered tests of achievement, cognitive ability, effortful control, and motivation and were interviewed with regard to their perceived competence and attitudes toward school. All children who spoke any Spanish (according to their teacher) or who were in bilingual classrooms or were classified as LEP were administered a test of English and Spanish proficiency by a bilingual examiner to determine the language in which the child was tested and the version of the achievement test administered. Achievement was tested with the Broad Reading and Broad Writing scales of the Woodcock-Johnson III Tests of Achievement (Woodcock, McGrew, & Mather, 2001) or the comparable Spanish achievement test, the Bateria-R (Woodcock & Muñoz-Sandoval, 1996).

**Texas Assessment of Knowledge and Skills: Reading and Math.** The TAKS measures a student’s mastery of the state-mandated curriculum, the Texas Essential Knowledge and Skills (TEKS). Both English and Spanish versions are available. The third grade TAKS includes both reading and mathematics tests. Students served by special education who meet certain eligibility requirements take an alternate assessment (TAKS Alternate) in reading, math, or both to determine student learning progress. TAKS Alternate is an alternate assessment based on alternate academic achievement standards and involves teachers observing students as they complete teacher-designed activities that link to the grade-level TEKS curriculum. Only students who took the standard TAKS are included in the current study. TEA establishes the percentage of correct answers that constitutes a “passing” rate for the reading and math tests (Texas Education Agency, 2008a). Information on the development of the TAKS Reading and Math is available from Texas Education Agency (2004). Promoted students took the third grade TAKS in spring 2004. Retained students took the third grade TAKS in spring 2005. Although students are given three opportunities to pass the TAKS, only scores from their first attempt were used in these analyses because only students who fail the first administration are given additional test administrations.

**Propensity Score Estimation and Analyses**

**Missing data.** Not all 769 participants had complete data on the 67 variables used to calculate propensity scores. Complete data were available for school records, including retention status. Other sources of data had some incomplete data. The percentage of data by other sources was 64% parent data, 68% sociometric data, 95% child test or interview data, and 86% teacher data. Not all students had scores on the third grade TAKS. Of the 769, 634 had TAKS reading scores and 643 had TAKS math scores. A student might be missing TAKS scores for a number of reasons, including lack of data from the school district for those students who had moved from one of the participating school districts or special education eligibility. Students served by special education may take an alternate test based on different standards, in which case TAKS scores
are missing. Although we attempted to obtain TAKS scores for students who moved outside one of the three participating districts, not all schools provided the requested data. Also, some students who moved out of the district were subsequently home-schooled or attended a private school (and therefore were exempt from TAKS testing). To determine if missing the TAKS was associated with retention status at grade 1, we performed χ² tests on the sample of 769 children. The tests yielded significant results (χ² (1) = 14.355, \( p < .001 \); χ² (1) = 12.050, \( p = .001 \)), indicating that retained children were more likely than promoted children to be missing the TAKS. Table 1 reports the number and percentage of students who were missing TAKS scores due to moving, special education placement, or other reasons. The “other” reason is based on records provided by the participating school districts and includes illness and cheating. Separate chi-square analyses tested whether retained and promoted students differed with respect to any of the three reasons for missing the TAKS reading or math scores. The groups did not differ on special education as the reason for missing for reading or math but did differ on moving as the reason for missing for reading (χ² (1) = 13.699, \( p < .001 \); χ² (1) = 15.278, \( p < .001 \)). More retained students missed TAKS scores due to moving compared with promoted students. In addition, the groups did not differ on “other” for math but did differ for reading (χ² (1) = 5.893, \( p < .015 \)). Retained students were more likely to miss TAKS reading scores due to other reasons.

**Creation of imputed datasets.** We created five multiple imputations of the dataset. We included all variables in the dataset in our imputation model. Because our intention was to fit a multi-level logistic model, we had to ensure that our outcome variables, TAKS passing status, and retention status (our outcome variable for the propensity score model) remained binary even after imputation. We achieved this by constraining imputed values to fall between 0 and 1 and rounding resulting imputed values. We used SAS PROC MI to conduct the multiple imputations and ensured that the Markov Chain Monte Carlo algorithm used by PROC MI properly converged by visually examining trace plots and auto-correlation plots.

Multiple imputation assumes a condition known as missing at random (MAR; Rubin, 1987). MAR assumes that missing data are related to observed variables in the dataset. As described above, children can have a missing TAKS score for many reasons. We believe that our exhaustive set of covariates captures most of the reasons that children are missing the TAKS. Even if unmeasured variables would affect missingness, we believe that the set of observed covariates would necessarily correlate with these unobserved causes, therefore lending credence to the missing at random assumption.
Propensity score analysis. After the five imputed datasets were created, we conducted a propensity score analysis within each of the imputed datasets. Our analytic strategy to deal with the missing data problem in conjunction with the propensity score analysis follows recommendations by Hill (2004). In particular, we (a) fitted propensity scores individually within each imputed dataset, (b) created five imputed and matched datasets based on the estimated propensity score, and then (c) analyzed each of the five sets individually using GEE models. In a final step, we combined estimates of our effects of interest following recommendations by Barnard and Rubin (1999) using a spreadsheet made available by von Hippel of Ohio State at www.sociology.ohio-state.edu/people/ptv/faq/MI_spreadsheet/MI_spreadsheet.htm.

Within each of the datasets, we estimated a propensity score using a logistic regression equation. The predictor terms were chosen based on an initial stepwise algorithm that retained all predictors that were significant at an alpha level of .20. After the estimation of the propensity score, we matched individual retained children with promoted children. Given the large reservoir of promoted children in comparison with the small amount of retained children, we allowed up to five promoted children to be matched with a retained child within each pair. Matching was conducted using nearest neighbor matching without replacement and a caliper width of .25 of a standard deviation of the estimated propensity score. The matching was performed using the R package MatchIt (Ho, Imai, King, & Stuart, 2007). For one to many matching schemes, the MatchIt package automatically creates weights that are based on the ratio of promoted to retained children for each match and then rescaled to sum to 1 to allow an estimation of the average effect of the treated. These weights were used in the analysis model to estimate treatment effects.

Creation of five matched datasets. Following this initial solution, we checked balance on all covariates by computing standardized differences, as suggested by Ho et al. (2007). Recall that the set of all covariates was larger than the set that was used to estimate the propensity score, because the initial set was chosen based on a stepwise regression algorithm. The initial model showed some large imbalances. As a result, we added additional terms to the propensity score model. We did this procedure of balance checking and re-estimating the propensity score with additional models several times, following the original recommendation of Rosenbaum and Rubin (1983), iteratively arriving at a solution that had very good overall balance in each of the five imputed datasets. Table 2 provides descriptive statistics for one of the five matched datasets.

The final propensity score model included a total of 67 variables. The appendix presents balance statistics that were averaged across all five imputed datasets (balance did not differ substantially across individual imputations). As evident in Table 3, balance was generally very good and a vast majority (93.2%) of covariates had standardized differences below 0.1. On average, only three variables had slightly remaining imbalances, however, the largest observed difference across all five imputed sets was only .188. Of special importance is that variables that were highly related to the outcome (e.g., Woodcock-Johnson scores) had excellent balance between the groups, eliminating any potential bias-inducing properties of these variables. Following recommendations of Schafer and Kang (2008), we included the estimated propensity score in the final analysis model, even though we matched on it, to control for any small remaining imbalances and to increase power of the statistical test.

Data Analysis

The average sample size across the five imputed datasets was 321 participants nested in 34 schools at Year 1. The GEE (Ghisletta & Spini, 2004; Hardin & Hilbe, 2002; Schafer, 2006) were estimated separately for the reading and math TAKS scores using HLM (V6.06; Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2004). The method of GEE is an extension of the generalized linear models, which can take correlated/clustered data with noncontinuous (e.g., binary or categorical) outcome variables into account and produce consistent and unbiased parameter estimates and corresponding standard errors regardless of the structure of the working correlation matrix (Ghisletta & Spini, 2004). We adopted the logit link in our analyses given that the
TAKS score had only two possible outcomes (i.e., 0 = fail, 1 = pass). GEE models were estimated for each outcome of the five imputed datasets, resulting in a total of 10 models. The setup of the GEE models was in the multi-level model framework as presented in equations 1 to 3. At level 1 (i.e., student level), we included both retention status at first grade (i.e., retent1\(_i\)) and the grand-mean centered propensity score of being retained (i.e., propensity\(_g\)) as predictors, and the outcome was the logarithm of the odds of passing the TAKS. The subscript \(i\) denotes students and subscript \(j\) denotes schools (e.g., retent1\(_g\) refers to the Year 1 retention status of the \(i\)th student nested within the \(j\)th school). Neither school-level random effects nor predictors were included in the level 2 (i.e., school level) model. The reason for not including any school-level random effects (i.e., \(u_{0j}\)) is that GEE is a population-averaged model and the setup of the GEE model does not include higher level (i.e., school level) random effects in the model. In addition, the variance of the school-level random effect was not significant. The combined model (i.e., substituting equation 2 back in equation 1) is presented in equation 3.

### Table 2

**Selected Descriptive Statistics for First Matched Sample (N = 339)**

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Retained</th>
<th>Promoted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>339</td>
<td>88 (26.0%)</td>
<td>251 (74%)</td>
</tr>
<tr>
<td>Categorical variables</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% male</td>
<td>59.3</td>
<td>58.0</td>
<td>59.8</td>
</tr>
<tr>
<td>% Caucasian</td>
<td>32.7</td>
<td>33.0</td>
<td>32.7</td>
</tr>
<tr>
<td>% bilingual placement</td>
<td>18.3</td>
<td>17.1</td>
<td>18.7</td>
</tr>
<tr>
<td>% high school or less</td>
<td>9.1</td>
<td>9.1</td>
<td>9.2</td>
</tr>
<tr>
<td>Continuous variables</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age at entering first grade</td>
<td>6.52(SD=0.39)</td>
<td>6.49(SD=0.30)</td>
<td>6.53(SD=0.41)</td>
</tr>
<tr>
<td>IQ</td>
<td>92.52(SD=14.25)</td>
<td>91.91(SD=14.75)</td>
<td>92.73(SD=14.09)</td>
</tr>
<tr>
<td>First grade WJ Reading</td>
<td>94.10(SD=17.45)</td>
<td>92.64(SD=18.35)</td>
<td>94.61(SD=17.13)</td>
</tr>
<tr>
<td>First grade WJ Math</td>
<td>100.56(SD=14.99)</td>
<td>100.79(SD=14.21)</td>
<td>100.48(SD=15.28)</td>
</tr>
<tr>
<td>Propensity score</td>
<td>.27(SD=.22)</td>
<td>.38(SD=.25)</td>
<td>.23(SD=.20)</td>
</tr>
</tbody>
</table>

**Note.** % high school or less is the percentage of students for whom highest adult education in home is high school degree or less; WJ Reading is Woodcock-Johnson III Broad Reading W score; WJ Math is Woodcock-Johnson III Broad Math W score.

### Table 3

**Frequency of Absolute Standardized Difference Scores on the 67 Covariates Between Retained and Promoted Students for First Matched Sample**

<table>
<thead>
<tr>
<th>Absolute standardized difference</th>
<th>Average count</th>
<th>Average %</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;.05</td>
<td>46</td>
<td>68.7%</td>
</tr>
<tr>
<td>.05 &lt; x &lt; .1</td>
<td>16.4</td>
<td>24.5%</td>
</tr>
<tr>
<td>.1 &lt; x &lt; .125</td>
<td>2.8</td>
<td>4.2%</td>
</tr>
<tr>
<td>.125 &lt; x &lt; .175</td>
<td>1.6</td>
<td>2.4%</td>
</tr>
<tr>
<td>&gt;.175</td>
<td>0.2</td>
<td>0.3%</td>
</tr>
</tbody>
</table>

**Note.** Balance statistics were nearly identical across the five datasets.
Retention and Performance on High Stakes Tests

**Results**

For the five imputed datasets, a statistically significant effect ($p < .05$) of retention was found in five datasets for the logit of passing the math TAKS and in four datasets for the logit of passing the reading TAKS. The parameter estimates and standard errors obtained from the two GEE models (i.e., passing reading TAKS and math TAKS) were averaged across the five model results according to methods suggested by Barnard and Rubin (1999). The averaged results are summarized in Table 4. The degree of freedom for the $t$ statistics is based on the average sample size ($N = 321$).

### TABLE 4

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Fixed effect</th>
<th>Coefficient (SE)</th>
<th>$t$ value</th>
<th>Odds ratio</th>
<th>Effect size$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>$\gamma_{00}$ (intercept)</td>
<td>1.41 (0.18)</td>
<td>7.93</td>
<td>4.08</td>
<td>2.25</td>
</tr>
<tr>
<td></td>
<td>$\gamma_{10}$ (propensity$_{ij}$)</td>
<td>$-0.58$ (0.69)</td>
<td>$-0.85$</td>
<td>0.56</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>$\gamma_{20}$ (retent$_{1ij}$)</td>
<td>0.59 (0.34)</td>
<td>1.75</td>
<td>1.80</td>
<td>0.99</td>
</tr>
<tr>
<td>Math</td>
<td>$\gamma_{00}$ (intercept)</td>
<td>1.50 (0.16)</td>
<td>9.62</td>
<td>4.48</td>
<td>2.48</td>
</tr>
<tr>
<td></td>
<td>$\gamma_{10}$ (propensity$_{ij}$)</td>
<td>$-1.29$ (0.40)</td>
<td>$-3.27$</td>
<td>0.27</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>$\gamma_{20}$ (retent$_{1ij}$)</td>
<td>0.57 (0.29)</td>
<td>1.95</td>
<td>1.76</td>
<td>0.97</td>
</tr>
</tbody>
</table>

$a$. Based on Chinn’s (2000) method, which converts odds ratio into Cohen’s $d$.

**Reading in third grade** ($\gamma_{10} = -0.58, t = -0.85$, $df = 318, p = .40$), when controlling for the influence of retention status. Retention status at grade 1 was marginally significantly related with passing the TAKS Reading at grade 3 ($\gamma_{20} = 0.59, t = 1.75, df = 318, p = .08$). The expected odds of passing the reading TAKS for retained students with an average propensity score was 1.80 times (i.e., $e^{0.59} = 1.80$) higher than their promoted counterparts who did not experience retention at grade 1. Based on Chinn’s (2000) method, we converted the odds ratio into Cohen’s $d$. The resulting effect size of retention on passing the Reading TAKS was substantial (.99), indicating practical significance of our results. Given the weighted average propensity score of 0.286 across five imputed datasets, the probability of passing the reading TAKS at grade 3 was 64.6% for retained students but only 55.3% for promoted students.

### Association Between Retention and Reading TAKS Passing

As shown in Table 4, the propensity score was not associated with passing the TAKS Reading in third grade ($\gamma_{10} = -0.58, t = -0.85$, $df = 318, p = .40$), when controlling for the influence of retention status. Retention status at grade 1 was marginally significantly related with passing the TAKS Reading at grade 3 ($\gamma_{20} = 0.59, t = 1.75, df = 318, p = .08$). The expected odds of passing the reading TAKS for retained students with an average propensity score was 1.80 times (i.e., $e^{0.59} = 1.80$) higher than their promoted counterparts who did not experience retention at grade 1. Based on Chinn’s (2000) method, we converted the odds ratio into Cohen’s $d$. The resulting effect size of retention on passing the Reading TAKS was substantial (.99), indicating practical significance of our results. Given the weighted average propensity score of 0.286 across five imputed datasets, the probability of passing the reading TAKS at grade 3 was 64.6% for retained students but only 55.3% for promoted students.

**TABLE 4**

*Average Generalized Estimating Equations (GEE) Fixed Effect Estimates for Reading and Math Texas Assessment of Knowledge and Skills (TAKS) Passing Rates for Five Imputed Datasets*

- **Reading**
  - $\gamma_{00}$ (intercept): 1.41 (0.18) $t = 7.93$, Odds ratio: 4.08, Effect size: 2.25
  - $\gamma_{10}$ (propensity$_{ij}$): $-0.58$ (0.69) $t = -0.85$, Odds ratio: 0.56, Effect size: 0.31
  - $\gamma_{20}$ (retent$_{1ij}$): 0.59 (0.34) $t = 1.75$, Odds ratio: 1.80, Effect size: 0.99
- **Math**
  - $\gamma_{00}$ (intercept): 1.50 (0.16) $t = 9.62$, Odds ratio: 4.48, Effect size: 2.48
  - $\gamma_{10}$ (propensity$_{ij}$): $-1.29$ (0.40) $t = -3.27$, Odds ratio: 0.27, Effect size: 0.15
  - $\gamma_{20}$ (retent$_{1ij}$): 0.57 (0.29) $t = 1.95$, Odds ratio: 1.76, Effect size: 0.97

*a. Based on Chinn’s (2000) method, which converts odds ratio into Cohen’s $d$.**
Association Between Retention and Math TAKS Passing

As shown in Table 4, the propensity score was negatively associated with passing the math TAKS in third grade ($\gamma_{10} = -1.29$, $t = -3.27$, $df = 318$, $p = .001$) after controlling for the influence of retention status. In other words, an increased probability of retention was associated with a decreased probability of passing the math TAKS. To meaningfully interpret the magnitude of the coefficient, we computed the change in odds associated with a one unit change in standard deviations. An increase of 1 standard deviation on the probability of retention was associated with a multiplicative change in odds of .742 (i.e., 1 standard deviation higher implies .742 times the original odds).

In addition, retention status at grade 1 was significantly related to passing the math TAKS at grade 3 ($\gamma_{20} = .57$, $t = 1.95$, $df = 318$, $p = .05$). The expected odds of passing the math TAKS for retained students was 1.76 times (i.e., $e^{.57} = 1.76$) higher than the promoted students. The effect of retention on passing the math TAKS was also large (.97) in terms of Cohen’s $d$. For students with an average propensity score, the probability of passing the TAKS at grade 3 was 62.9% for the retained students and only 53.0% for the promoted students.

Discussion

To the best of our knowledge, this research presents the first study to investigate the association between early grade retention and subsequent performance on a statewide, curriculum-aligned test of reading or math. We used propensity score matching to address potential selection bias due to pre-existing differences between retained and promoted students. Our propensity score model included a large number of variables theoretically and empirically associated with both selection for grade retention and achievement. Balance diagnostics revealed that our propensity score matching achieved very good balance between retained and promoted students on a comprehensive set of 67 baseline variables. By including the propensity score as an additional covariate in the GEE analyses, we provide additional assurance that the results from the GEE models provide an accurate estimate of the effect of treatment on TAKS passing scores. In summary, the combination of (1) a low achieving sample, (2) use of propensity scores to obtain a matched sample of retained and promoted students balanced on relevant baseline variables, and (3) use of propensity scores as a covariate greatly reduces the probability that estimated treatment effects are biased due to selection. To increase the likelihood of model convergence, we decided to reduce the complexity of the model by excluding the matched-group level (given the nonsignificant variance) and analyzed the data using the GEE model as presented in equations 1 to 3 with students nested within schools.

Analyses were then conducted using GEE, which can produce consistent and unbiased parameter estimates and corresponding standard errors while taking school-level effects into account. Averaged across the five multiply imputed and matched samples, students retained in grade 1 were 1.80 times higher on the odds of passing the reading TAKS at grade 3 than students promoted in grade 1. In other words, retained (at grade 1) students were more likely than promoted students to receive a passing score. For students with an average propensity score, the probability of passing the reading TAKS at grade 3 was 64.6% for the retained students and only 55.3% for the promoted students. Similar results were obtained for the odds of passing the TAKS math test. Retained students were 1.76 times higher on the odds of passing the math TAKS in grade 3 than their promoted counterparts. That is, retained students were more likely to receive a passing score on the math TAKS at grade 3 than the promoted students. For students with an average propensity score, the probability of passing the TAKS at grade 3 was 62.9%, but only 53.0% for a promoted student.

Comparing These Results With Prior Research

The finding of a positive association between grade retention and achievement is counter to the corpus of literature on retention effects. Jimerson (2001a) found that of 175 achievement effect sizes for achievement outcomes reported in 20 studies, 84 (52%) reported no statistically significant effect for retention. Of the remaining 91 effects, only 9 favored the retained students, whereas 82 favored the comparison group of promoted students. There are several possible reasons for discrepant results between this study and
the majority of previously published studies on the effects of retention. First, this is one of the few studies to use propensity score methods to remove baseline differences on a wide range of relevant child, family, peer, and school variables. When properly used, propensity scores have shown to substantially reduce or eliminate selection bias in observational studies and strengthen causal reasoning (Rosenbaum & Rubin, 1983; Shadish, Luellen, & Clark, 2006). By balancing retained and promoted students on these vulnerability factors, the opportunity for unmeasured vulnerabilities that are more characteristic of grade retained students to affect results is greatly minimized. This reasoning is consistent with recent findings of a meta-analysis of 22 studies published between 1990 and 2007 that employed comparison groups and varying levels of control for baseline vulnerabilities between retained and promoted students (Allen et al., 2009). Studies employing better controls (i.e., higher design quality) produced less negative results than did studies employing poorer controls. Furthermore, the effect size for high quality designs was not statistically significantly different from zero.

A second reason for results that diverge from previous studies is that the current study employed a curriculum-aligned measure of academic achievement administered to all students during their 1st year in third grade. Thus, these scores represent same-grade versus same-age comparisons. Same-grade comparisons evaluate retained and promoted students when they are in the same school grade. Thus, retained students are assessed 1 year later than are their promoted peers. In contrast, same-age comparisons assess students at the same age, or calendar year, usually on an age-standardized measure of achievement. Same-grade comparisons are more likely to find benefits of retention, at least in the short term, than are same-age comparisons (Wu et al., 2008b). The more favorable result for same-grade comparisons may be more pronounced when the achievement measure is aligned with the curriculum. Given the consequences for students, teachers, and schools associated with performance on the TAKS, it is likely that teachers focus instruction on those competencies or “essential elements” tested with the TAKS. By the time retained students took the third grade TAKS, they had been exposed to the curriculum on which the test is aligned for 4 years, whereas promoted students had been exposed to the curriculum for only 3 years. The higher passing rate of retained students may reflect this difference in intervention “dosage.”

Third, the standards-based reform movement and its associated emphasis on high stakes testing and ending social promotion may account for differences between the current study’s findings and previous research on grade retention. That is, retention effects may differ in different educational policy contexts. One objective of the use of tests for making decisions with regard to promotion is to standardize and objectify grade retention decisions and move away from social promotion (U.S. Department of Education, 1999). Previous published research on retention suggests that prior to the adoption of policies requiring the use of tests for making retention decisions, a number of nonachievement-related variables entered into decisions to retain a student (Jimerson, Carlson, Rotert, Egeland, & Sroufe, 1997; McCoy & Reynolds, 1999). A study on predictors of grade retention for this longitudinal sample (Willson & Hughes, 2009) found that many demographic variables (e.g., economic disadvantage, social, emotional, and behavioral functioning) that predicted achievement in previous studies did not make a unique contribution to the prediction of retention in first grade, above measures of students’ academic competencies. A change in criteria used to select students for the retention intervention may be associated with a change in the effectiveness of the intervention.

It is also possible that in the current educational policy environment, academically at-risk students in the early grades receive more early intervention services than was true in the past. Since 1999–2000, school districts have provided accelerated instruction to students identified as at risk for reading or mathematics difficulties (Texas Education Agency, 2007). However, data on the nature, extent, and quality of these services and the basis for selecting students for accelerated instruction are not available.

Selection of Achievement Standard in Retention Research

It is instructive to compare these results with those reported in a companion published study using an overlapping sample. Wu et al. (2008b) used propensity matched pairs and two-piece
linear growth curve models to examine the short-term and longer term effects of grade retention on change of Woodcock-Johnson III or the comparable Spanish language version (Woodcock et al., 2001; Woodcock & Muñoz-Sandoval, 1996) math and reading age and grade scores across 4 years. Results differed based on scale used (age or grade equivalent scores), time elapsed since retention year, and achievement domain (reading vs. math). For reading, when grade equivalent scores were used (similar to same-grade comparisons), retained students increased in reading more than promoted students did in the short term (i.e., when retained students were repeating the grade) but grew more slowly in the longer term. When age-based (i.e., Rasch-based W scores) scores were used, retained students grew more slowly in the short term but more rapidly in the longer term, relative to promoted peers, such that by Year 4, retained and promoted students did not differ in reading. The current study adds to the findings of Wu et al. by comparing retained and promoted students’ performance on a curriculum-aligned measure of reading and math taken by all students during their 1st year in third grade. On this same-grade comparison, retained students outperform their propensity-matched promoted peers.

Given that the effects of retention differ based on the standard used (age or grade; curriculum-aligned test or nationally standardized test), the choice of the standard by which to answer the question, “Is grade retention an efficacious intervention?” demands greater consideration in the literature. The Allen et al. (2009) study is the first to specifically investigate the effect of grade versus age comparison on retention outcomes. They found that effect sizes for retention were more positive (or less negative) when grade versus age comparisons were made and that this difference decreased with increasing years post-retention. That is, with additional years post-retention, the relative benefit of grade versus age comparisons declined.

On a practical level, schools view a passing score on the TAKS as a strongly positive outcome for the school and the district. Although grade-level tests must be administered to students beginning in kindergarten, results of schools’ performance do not enter into the state education accountability system (Academic Excellence Indicator System [AEIS]) until third grade. AEIS performance ratings range from academically unacceptable to exemplary. These ratings, as well as extensive information on a number of performance indicators, including TAKS passing rate disaggregated by ethnicity, sex, special education, low income status, and LEP status, for every public school and district, are easily accessible via the Texas Education Agency website (http://ritter.tea.state.tx.us/perfreport/aeis/). The percentage of students passing the TAKS, especially in historically low achieving subgroups, is the primary determinant of a school or district’s performance ratings (Texas Education Agency, 2008b). The Texas Education Agency allocates penalties and rewards based on these ratings. Perhaps more consequential, ratings are widely publicized in print and broadcast media and are sent to parents of all public school students. TAKS scores enter into the evaluation of teachers and administrators. In response to these consequences, schools and teachers engage in a variety of “triage” strategies to increase the passing rate, including allocating disproportionate resources to “bubble” students who are just below the passing score and to “accountable” students according to the performance rating system (Booher-Jennings, 2005). Similar motivating effects of high stakes testing on educational practice have been documented in other states (Achieve, 2002; Watanabe, 2007).

Strength, Limitations, and Future Research

Study findings must be interpreted in the context of the study’s strengths and limitations. Counted among the strengths of this investigation is use of propensity matching to create samples of retained and promoted students who were balanced on a large number of academically relevant variables at baseline, combined with use of the propensity score as a covariate in the GEE analyses. Such procedures greatly reduce the possibility that any differences in performance on the third grade TAKS were due to pre-retention differences between retained and promoted students. In addition, because prior empirical evidence has found school-level effects in retention rates (Hong & Raudenbush, 2005; Hong & Yu, 2008), use of GEE provides a more accurate estimate of the effect of the retention decision, above the effect of the school in which the child was enrolled in first grade. Finally, the use of
passing performance on the third grade state accountability test rather than on measures most often used in previous research on the effects of retention (i.e., performance on a nationally standardized test of achievement) increases the applied significance of the findings to educators and educational policy makers.

One limitation of the study is the amount of missing data on TAKS performance and group differences in level of missingness. Retained students were more likely than promoted students to be missing TAKS reading (26.7% for retained vs. 15.1% for promoted) and math scores (26.1% for retained vs. 13.7% for promoted). An analysis of reasons that students were missing TAKS scores revealed that retained students were more likely to be missing TAKS reading and math scores due to having moved from the district but not due to special education eligibility. Relative to promoted students, retained students were more likely to be missing TAKS math (but not reading) for the “other” reason code.

The greater mobility among retained students may be due to the fact that they took the third grade TAKS 1 year later than did their promoted peers. Thus, they had 1 more year in which to move from one of the participating school districts. It is also possible that retention itself increases the probability that a student leaves the district, either to be home-schooled, to attend private school, or to get a “fresh” start in a new public school. Although we do not know the reason for greater mobility for retained students, we do know that retained and promoted students did not differ in special educational placement. This finding is important because special education placement is considered a type of educational failure, and students “selected” for grade retention closely resemble students selected for special education placement (Beebe-Frankenberger, Bocian, MacMillan, & Gresham, 2004). Because the “other” category is a heterogeneous category and occurs relatively infrequently (N = 12 for reading and N = 12 for math), we do not know why retained students were more likely to be missing the TAKS due to this reason category.

Our data imputation procedures minimize concerns that differences between retained and promoted students on missing TAKS affect the GEE results. If missingness is related to unobserved outcomes or the outcome itself (a situation referred to as missing not at random), internal validity of results is threatened. However, if missingness is related to observed covariates, an assumption called missing at random holds. Although there is no statistical test for missing at random, for reasons described above, we strongly believe that this assumption holds in our dataset. We have a rich set of covariates from which missing values can be reasonably imputed. Under missing at random, multiple imputation yields unbiased results. This assumes no unobserved variables that have large unique effects on missingness and a correct specification of the functional form relating covariates and missingness. The second problem for causal inference is the selection bias due to nonrandomized assignment to retention or promotion. We tried to address this problem with the use of propensity scores. The assumptions that we make are that we have included all important covariates and have no unobserved variables that bias results, have properly specified the functional form of the propensity score model, and have sufficient overlap between the groups. We believe that we have addressed both problems to causal inference.

As is always the case, additional research is needed to better understand these results. For example, it would be instructive to directly compare effects of retention on performance on high stakes tests that are aligned with the curriculum with the effects of retention on nationally standardized tests of reading and math, using same-grade comparisons. Because effects of retention differ based on number of years post-retention, it will be important to investigate whether the benefit of early grade retention on state accountability tests is maintained as students continue in school and whether early retained and promoted students differ on more distal measures of achievement, such as school completion. It is also important to evaluate the effect of grade retention within different policy contexts. In the context of accountability testing on measures of minimum grade-level competencies, retention may have a different effect than it would in a different policy context.

**Conclusion**

It is important to acknowledge that associations between grade retention and achievement
Hughes et al.

differ based on the standard used. This study used statewide criterion-referenced tests based on statewide curriculum standards for third grade students. Our results suggest, but cannot prove, that students who are retained in first grade are more likely to pass these tests than they would have been if they had been promoted to second grade. If the purpose of retention is to be a one-time adjustment in the student’s academic pathway, these results suggest that retention served its intended purpose. These findings challenge the conventional wisdom (a wisdom informed by previous empirical research) that grade retention is harmful to students’ future academic performance. In the context of high stakes testing and the establishment of promotional gates for grade retention, retention in the early grades may increase a student’s chance of successfully meeting the academic challenges of subsequent grades.

Note

1. The standard deviation of the propensity score was .23, which implies change on the odds ratio of \((.23) \times (-1.29) = -.297\), which can be transformed to odds of \(e^{-0.297} = .742\).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

Financial Disclosure/Funding

This research was supported in part by a grant to Jan N. Hughes from the National Institute of Child Health and Human Development (5 R01 HD39367).

References


Ho, D., Imai, K., King, G., & Stuart, E. (2007). Matching as nonparametric preprocessing for...


Hughes et al.


Authors

JAN N. HUGHES, Ph.D., is a professor of educational psychology at Texas A&M University. Her current research interests include reciprocal effects between students’ relationships with teachers and peers and students’ behavioral and academic adjustment, conduct disorders, and school psychology.

QI CHEN, Ph.D., is an assistant professor in the Educational Psychology Department at the University of North Texas. Her current research interests include longitudinal data analysis with a special emphasis on latent class trajectory analysis and multi-level modeling (MLM) techniques.

FELIX THOEMMES is an assistant professor in the Department of Educational Psychology at Texas A&M University. His area of research is causal inference in non-experimental studies and the use of propensity scores. He is currently a visiting scholar at the University of Jena, Germany.

OI-MAN KWOK, Ph.D., is an assistant professor of educational psychology at Texas A&M University. His current interests include longitudinal data analysis using both structural equation modeling (SEM) and multi-level modeling (MLM) techniques.