ELSEVIER

Journal of School Psychology
46 (2008) 85–105

Journal of School Psychology

Short-term effects of grade retention on the growth rate of Woodcock–Johnson III broad math and reading scores ☆

Wei Wu [a,*], Stephen G. West [a], Jan N. Hughes [b,*]

[a] *Department of Psychology, Arizona State University, Tempe, AZ 85287-1104, USA*
[b] *Department of Educational Psychology, Texas A&M University, College Station, TX 77843-4225, USA*

Received 27 June 2006; received in revised form 19 January 2007; accepted 26 January 2007

Abstract

We investigated the effects of grade retention in first grade on the growth of the Woodcock–Johnson broad mathematics and reading scores over three years using linear growth curve modeling on an academically at-risk sample. A large sample ($n=784$) of first grade children who were at risk for retention was initially identified based on low literacy scores. Scores representing propensity for retention were constructed based on 72 variables collected in comprehensive baseline testing in first grade. We closely matched 97 pairs of retained and promoted children based on their propensity scores using optimal matching procedures. This procedure adjusted for baseline differences between the retained and promoted children. We found that grade retention decreased the growth rate of mathematical skills but had no significant effect on reading skills. In addition, several potential moderators of the effect of retention on growth of mathematical and reading skills were identified including limited English language proficiency and children's conduct problems.

Keywords: Grade retention; Linear growth curve modeling; Propensity score; Optimal matching

☆ This research was supported in part by grant to Jan Hughes from the National Institute of Child Health and Development (5 R01 HD39367-02). This article was accepted under Dr. Pianta's editorship.
* Corresponding authors.
E-mail addresses: wei.wu.1@asu.edu (W. Wu), sgwest@asu.edu (S.G. West), jhughes@tamu.edu (J.N. Hughes).

0022-4405/$ - see front matter © 2007 Society for the Study of School Psychology. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.jsp.2007.01.003

Children who do not meet minimum academic competencies create a dilemma for the educational system. Historically, parents and teachers needed to decide whether these children should be promoted to the next grade in the hope they would somehow acquire the necessary competencies ("social promotion"), or be retained in grade ("grade retention") to give the children another opportunity to reach the appropriate mastery level. Although the prevalence of grade retention as an educational practice has shown large shifts across the past 40 years, it remains a common educational practice. In 2004, US Census data revealed that 9.6% of US youth ages 16–19 had been retained in grade one or more times. Recently nine states have initiated legislation that requires students to pass tests of grade level competencies in order to be promoted to the next grade. The increase beginning in the mid-1990s in establishing promotional gates, a movement consistent with No Child Left Behind's emphasis on accountability testing at earlier grades, suggests the pendulum is again moving toward higher rates of grade retention (Bali, Anagnostopoulos, & Roberts, 2005).

Statistics on grade retention that are based on all youth mask large geographical, gender, and ethnic/racial differences in the use of this educational intervention. In 2004, youth in the South were nearly three times as likely to have ever been retained in a grade (14.0%) as were youth in the West (5.4%). Males were more than twice as likely to be retained as females. Whereas 16.1% of Black youth were ever retained in grade, 8.1% of White and 9.2% of Hispanic youth were ever retained in grade (National Center for Educational Statistics, 2006).

Retaining a child is an expensive educational practice. Based on the average per-pupil expenditure of $8916 for the 2004–2005 academic year, Texas spent an estimated 1.7 billion dollars for the extra year of schooling for the 190,802 children retained in grades K-12 in the 2000–2001 academic year (Texas Education Agency, 2006a).

Effects of grade retention

Laws and public policy on grade retention appear inconsistent with empirical research. Although some studies report short-term positive effects of grade retention on achievement and behavior (Karweit, 1999; Peterson, DeGracie, & Ayabe, 1987; Pierson & Connell, 1992), better designed studies that incorporate some control for pre-retention differences document a range of negative effects (Meisels & Liaw, 1993; Pagani, Tremblay, Vitaro, Boulerice, & McDuff, 2001; Reynolds, 1992). Importantly, the strong association between repeating a grade and dropping out of school before graduation remains after statistical adjustments for measured pre-retention differences, typically demographics, are made (Alexander, Entwisle, & Dauber, 2003; Grissom & Shepard, 1989; Holmes, 1989; Peterson et al., 1987; Reynolds, 1992). Jimerson's (2001) meta-analysis of 20 studies published in the 1990s concluded that retention has generally negative average effects on both academic and psychosocial functioning. Recent studies using growth curve modeling find that the effects of retention in the early grades on long-term adjustment either do not differ from those at later grades (Silberglitt, Jimerson, Burns, & Appleton, 2006) or are more negative (Pagani et al., 2001). These findings challenge the conventional wisdom that early grade retention is a less harmful "preventive" intervention than is retention in later grades (Tomchin & Impara, 1992).

Methodological limitations of previous research on effects of retention

Control for pre-existing differences

Although the weight of existing evidence indicates harmful effects of retention on academic and psychosocial outcomes, methodological limitations of extant research limit confidence in this conclusion. The most serious limitation is a failure to adequately control for a comprehensive set of pre-existing differences between retained and promoted students that affect students' academic and social trajectories, leaving open the possibility that grade retention is a marker of pre-existing vulnerabilities. With few exceptions (e.g., Alexander et al., 2003; Reynolds, 1992), studies have primarily controlled for demographic variables and have not provided adequate adjustment for ability-related factors and psychosocial adjustment factors associated with school performance.

Individual differences in response to retention

A second limitation of the existing research base is the failure to investigate factors that may moderate the effect of grade retention. Although inadequate mastery of grade level competencies is the predominant factor in most decisions to retain a child, a plethora of non-achievement related factors also come into play (Dauber, Alexander, & Entwisle, 1993; Jimerson, Carlson, Rotert, Egeland, & Sroufe, 1997; Mantzicopoulos, 2003; McCoy & Reynolds, 1999; Willson & Hughes, 2006). Presumably, non-achievement related factors associated with retention are based on educators' personal theories of which children are likely to benefit the most from repeating a grade. For example, the association between being young for grade and being retained (Mantzicopoulos, 2003; Willson & Hughes, 2006) may reflect the view that younger students are more likely to benefit from the "gift" of an extra year. However, the research basis for making decisions based on these factors is virtually non-existent. The few studies that have investigated moderating factors have usually restricted their analysis to distal demographic variables such as gender, family income, or parental educational level rather than more proximal variables such as child psychosocial adjustment (Meisels & Liaw, 1993; Pagani et al., 2001; Reynolds, 1992).

Study purpose

The present study extends past research by employing a prospective design with an ethnically and linguistically diverse sample. The children are never-before retained first grade children at risk for retention based on low literacy scores at the beginning of first grade. The study includes both small city and urban school districts with ethnically diverse populations. This study is the first investigation of grade retention effects to measure a comprehensive set of baseline variables and to use modern propensity score methods and optimal matching procedures (Rosenbaum, 2002; West, Biesanz, & Pitts, 2000). This procedure has great strength in identifying comparable retained and non-retained children and adjusting for a wide variety of baseline differences. Importantly, the propensity scores, which represent the probability of being retained in first grade, are based on 72 variables

assessed prior to any study children being retained in grade. Details of the propensity score estimation and matching procedures are included below. The present study focuses on the effects of retention on the estimated linear rate of growth of two cognitive abilities, reading and math, as measured by the Woodcock–Johnson III Tests of Achievement (Woodcock, McGrew, & Mather, 2001) or the comparable Spanish achievement test, the Batería-R (Woodcock & Muñoz-Sandoval, 1996).

We also investigate potential moderators of the effects of retention. Consistent with developmental systems theory (Lerner, 1989), we expect the impact of retention on a child's achievement trajectory will differ based on the interplay of the retention "treatment" and multiple factors both within and outside the developing child (Cicchetti & Posner, 2005; Sameroff, 1989). Thus we investigate whether the effects of retention differ based on two selected child socio-demographic characteristics (age; status as limited English proficient) and seven indices of psychosocial adjustment (parent-, and teacher-rated conduct problems and hyperactivity, peer-rated externalizing behaviors, quality of the teacher–student relationship quality, and personality resilience).

Socio-demographic variables

Several studies have documented associations between socio-demographic variables and retention (Beebe-Frankenberger, Bocain, MacMillan, & Gresham, 2004; Meisels & Liaw, 1993). However, studies utilizing statistical controls for psychosocial adjustment and achievement often fail to find independent effects for socio-demographic variables on retention (Dauber et al., 1993; Jimerson et al., 1997; Mantzicopoulos, 2003; McCoy & Reynolds, 1999; Willson & Hughes, 2006). The Chicago Public Schools longitudinal study (Reynolds, 1992) is one of the few studies investigating whether socio-demographic variables moderate the effects of retention. Reynolds found no evidence that the effects of retention in grades K-3 varied by any of the five socio-demographic variables investigated (i.e., gender, preschool participation, parent education, school mobility, and parent involvement).

In this study, we selected age at entrance to first grade and Limited English Proficient (LEP) status as potential socio-demographic moderators of the effect of retention. Age was selected based on evidence that being young in kindergarten makes an independent contribution to the prediction of retention above the effect of children's academic achievement and behavioral adjustment (Mantzicopoulos, 2003). Furthermore, in a study using all participants in the longitudinal study from which the current sample was selected (Willson, Hughes, Gleason, & Wong, submitted for publication), being young at entrance to first grade was the only independent socio-demographic predictor of grade retention, after controlling for a number of cognitive and achievement variables. We reasoned that if being young for grade increases a child's probability of being retained, it is important to determine if younger children respond more favorably to grade retention, relative to older children. If younger children do not benefit more from retention than do older children, the wisdom of considering age in retention decisions is called into question. We expected younger children would benefit more from retention than older children due to the greater advantages younger children would accrue from an extra year of cognitive, linguistic, and physical maturation, relative to their grade mates.

We selected LEP status as a potential moderator of retention effects based on two considerations. First, LEP students are a rapidly increasing but under-studied segment of elementary students in the United States. Between 1979 and 2004, the number of school-age children who spoke a language other than English at home increased from 3.8 to 9.9 million, or from 9% to 19% of all children in this age group, with Spanish speakers representing the bulk of the increase (National Center for Education Statistics, 2006). Second, Reynolds (1992) found that Hispanic and LEP students fared worse following retention on reading and perceived competence, but he cautioned that these results may be unstable due to the small number of Hispanic/LEP retainees ($N=10$) in the sample. It is important to attempt to replicate this finding in a somewhat larger sample of Hispanic LEP students.

Psychosocial adjustment

Very little research investigates whether children's social and behavioral adjustment in the classroom moderates the effects of retention; indeed, we were unable to locate any research that investigated the moderating role of children's relationships with their teachers, regulatory control, or adaptive personality traits on the effects of retention. We expected that the negative impact of grade retention would be stronger for children with good self-regulation, adaptive personalities, and positive relatedness to their teachers. Our reasoning is based on the assumption that students who are promoted to the next grade, relative to matched children who remained in the same grade a second year, will encounter more novel academic and social challenges. Also, promoted students do not have the advantage of an extra year of maturity, relative to their same-age classmates, whereas retained students will be, on average, a year older than their classmates. We expect that students with good self regulation (indexed by low levels of externalizing problems and high personality resilience) and positive relatedness to teachers are more likely to "rise to the occasion" when promoted; consequently, these children are expected to achieve more, relative to students with good self regulation, positive relatedness to teachers, and personality resilience who *repeat* first grade. Low-achieving children who are promoted and who lack these personal and social resources are expected to experience greater difficulty in coping with new and unfamiliar academic content and behavioral expectations.

The current study investigates the impact of being retained in first grade on children's academic achievement trajectories over the course of three academic years. Because failure to master grade level academic competencies is the primary factor in decisions to require a child to repeat a grade (Alexander, Entwisle, & Dauber, 1993; Beebe-Frankenberger et al., 2004; Jimerson et al., 1997; Mantzicopoulos, 2003), it is reasonable that the success of this educational intervention be gauged in terms of its impact on academic achievement.

Method

Participants

Our strategy was to select an initial sample of children at high risk for retention and then to identify a subsample of retained and non-retained children within the larger

sample for the growth curve analysis that were closely matched on a comprehensive set of baseline variables. The initial sample of participants who are participating in our longitudinal study were recruited from three school districts in Texas (1 urban and 2 small city) across two sequential cohorts in first-grade during the fall of 2001 and 2002. 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 1374 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 1200 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 ($n=784$) sample was 37% Hispanic, (39% of whom were Spanish language dominant), 34% White 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 and McCallum, 1998) for the sample was 92.91 (SD=18.01), and the mean reading achievement score was 96.40 (SD=14.28).

Participants for the growth curve analyses were those 196 children (58% male) who were successfully matched with respect to their propensity to be retained in first grade (see description of propensity matching procedures below) and who had achievement test scores for at least one of the three assessment periods. Otherwise stated, we were able to successfully identify a closely matched promoted child for 97 of the 165 children who were retained in first grade. The racial/ethnic composition of the sample was 33% Caucasian, 33% Hispanic, 31% African American, and 3% Asian/Pacific Islander. At entrance to first grade, children's mean age was 6.45 (SD=.33) years.

Design overview

Demographic information including child age and status as Limited English Proficient (LEP) were obtained from school district records. Teacher, parent, and peer data were collected in the Spring of Year 1, when all participants were in first grade. Teachers and parents received $25.00 for completing and returning the questionnaires. Peers' perceptions of the level of externalizing behaviors were obtained via individual interviews conducted between February and May of Year 1. Beginning in Year 1, annual measures of math and reading achievement were individually administered at school for three years, with the constraint that at least 8 months separated each annual assessment.

Measures

The following measures were given at each wave of measurement. Time 1 measures collected prior to retention were used in the estimation of propensity scores. A complete list of the 72 variables used in calculation of propensity scores is available from the third author. The measures of academic achievement served as the primary outcome measures for this study. Nine time 1 measures that might affect the rate of growth of children in response to retention versus promotion were explored as potential moderator variables. These measures and the achievement measures are described below.

Academic achievement

The *WJ-III Tests of Achievement* (Woodcock et al., 2001) is an individually administered measure of academic achievement for individuals ages 2 to adulthood. The WJ-III Broad Reading W Scores (Letter–Word Identification, Reading Fluency, Passage Comprehension subtests) and the WJ-III Broad Math W Scores (Calculations, Math Fluency, and Math Calculation Skills subtests) were used. The Reading and Math W scores are based on the Rasch measurement model, yielding an equal interval scale, which facilitates modeling growth in the underlying latent achievement (Khoo, West, Wu, & Kwok, 2006). Extensive research documents the reliability and construct validity of the WJ-III and its predecessors (Woodcock & Johnson, 1989; Woodcock et al., 2001).

The *Batería Woodcock–Muñoz: Pruebas de aprovechamiento — Revisada* (Woodcock & Muñoz-Sandoval, 1996) is the comparable Spanish version of the *Woodcock–Johnson Tests of Achievement — Revised* (*WJ-R*; Woodcock & Johnson, 1989), the precursor of the *WJ-III.* If children or their parents spoke any Spanish, children were administered the *Woodcock–Muñoz Language Test* (Woodcock & Muñoz-Sandoval, 1993) to determine the child's language proficiency in English and Spanish and selection of either the *WJ-III* or the *Batería-R*. The Woodcock Compuscore (Woodcock & Muñoz-Sandoval, 2001) program yields W scores for the *Batería-R* that are comparable to W scores on the *WJ-R*. The Broad Reading and Broad Mathematics W Scores were used in this study and are referred to as WJ reading and math scores.

Teacher and parent report of conduct problems and hyperactivity

Teachers and parents completed the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997), a brief (25-item) screening measure for psychopathology. Each item is rated on a 0–2 scale (i.e., not true, somewhat true, certainly true). The SDQ yields five scales comprised of 5 items each. The Conduct Problems and the Hyperactivity scales assess externalizing behaviors. For our sample, the coefficient alpha for Conduct Problems was .84 for teachers and .71 for parents. For Hyperactivity the coefficient alpha was .89 and .81 for teachers and parents, respectively. In a sample of children participating in this longitudinal study, teacher reports of Conduct Problems and of Hyperactivity are moderately to strongly correlated with both parent reports (.47 and .30, for conduct problems and hyperactivity, respectively) and with peer reports (.50 and .46, for conduct problems and hyperactivity, respectively) (Hill & Hughes, in press). Exploratory and confirmatory factor analyses support the construct validity of the teacher and parent versions of the SDQ (Dickey & Blumberg, 2004; Goodman, 2001; Hill & Hughes, in press).

Peer nomination of externalizing problems

Peers' perceptions of classmates' hyperactivity and aggression were obtained following procedures widely recommended in the peer assessment literature (Cillessen & Bukowski, 2000). In individual interviews, children were presented a roster with the names of all classmates. The interviewer read all classmates' names and asked the child if he or she knew each child. Then the interviewer asked the child to nominate as few or as many classmates as they wished who fit each descriptor. Of interest to this study are the aggression item ("Some kids start fights, say mean things, or hit others.") and the hyperactivity item ("Some kids do strange things and make a lot of noise. They bother people who are trying to work."). Each class member received an aggression and hyperactivity score based on the number of nominations that child received. Sociometric scores were standardized within classrooms. Because the two scores were highly correlated ($r = .74$), a composite peer-rated externalizing score was computed as the mean standardized score on the aggression and hyperactivity items. Written parent consent was obtained for each child who participated in the sociometric interview. However, all children in a classroom were eligible to be rated or nominated. Terry (1999) reported that reliable and valid sociometric data can be collected using the unlimited nomination approach when as few as 40% of children in a classroom participate. Thus, sociometric scores were computed only for children located in classrooms in which more than 40% of classmates participated in the sociometric assessment. The mean rate of classmate participation in the sociometric administrations was .65 (range .40 to .95). Elementary children's peer nomination scores derived from procedures similar to those used in this study have been found to be stable over periods from six weeks to four years and to be associated with concurrent and future behavior and adjustment (for review, see Hughes, 1990).

Resilient personality

A confirmatory factor analysis (Kwok, Hughes, & Luo, 2006) on a sample of 445 first grade children participating in the current longitudinal study supported a second order measurement model of resilient personality defined by three first order factors: agreeableness (9 items), conscientiousness (8 items), and ego resiliency (7 items). Agreeableness and conscientiousness items were taken from the scales of the same name of the Big Five Inventory (John & Srivastava, 1999). Sample agreeableness items include "is helpful and unselfish with others;" "likes to cooperate with others;" and "is sometimes rude to others" (reverse scored). Sample conscientiousness items include "does a thorough job;" "is a reliable worker;" and "tends to be disorganized" (reverse scored). Coefficient alpha for each scale was .94. Ego resiliency items were derived from items on the California Child Q Sort (Caspi, Block, Block, & Klopp, 1992). Sample items include "resourceful in initiating necessary activities" and "falls to pieces under stress" (reverse scored). Teachers responded to each of the items using a 5-point scale. Coefficient alpha for our sample was .85. A score for resilient personality was computed as the mean of the standardized score for each scale.

Teacher–student relationship quality

The 22-item Teacher Student Relationship Inventory (TSRI; Hughes, Cavell, & Willson, 2001) is based on the Network of Relationships Inventory (Buhrmester & Furman, 1987). Teachers indicated on a 5-point Likert-type scale their level of support (16 items, coefficient

alpha = .94) or conflict (6 items, coefficient alpha = .92) in their relationships with individual students. Because the support and conflict scales were negatively correlated (−.57), we recoded the conflict items in the opposite direction. The coefficient alpha of the 22 items after recoding the conflict items was .95. A total relationship quality score was computed as the mean of the 22 item scores. In a longitudinal study of behaviorally at-risk elementary students, the TSRI Support score predicted changes in behavioral adjustment and peer relationships (Meehan, Hughes, & Cavell, 2003). In the larger sample, the TSRI predicted cross-year changes in children's achievement (Hughes & Kwok, 2006).

Demographic variables

Information on age and designation as Limited English Proficient was obtained from school records.

Propensity score estimation

Propensity scores, the predicted probability of being retained in first grade, were estimated for 769 children for whom retention information was available using 72 background variables collected at the initial testing, including child demographic variables, child, peer, teacher, and parent data covering the areas of academic aptitude (e.g., the Universal Nonverbal Intelligence Test), academic achievement (Woodcock Johnson III or the Spanish-language Batería-R broad math and reading), personality (e.g., agreeableness; effortful control), behavioral and social adjustment, peer relations, and family adversity. Methods based on logistic regression (Rosenbaum, 2002; Rosenbaum & Rubin, 1983) were used to estimate propensity scores. The larger the propensity score, the larger the predicted probability that the child would be retained in the first grade. For the 165 retained children, the propensity score ranged from .0034 to .9892 with mean = .5402, SD = .2923. For the 504 promoted children, the propensity score ranged from .0003 to .9179 with mean = .1258 and SD = .1632. Note that 57.6% of the retained children had propensity score larger than .50, while only 5.5% of the promoted children had propensity scores larger than .50. The distribution of propensity scores prior to matching is shown in the top two panels of Fig. 1. Despite our selection of an at risk sample, these distributions indicate the need for close matching of the retained and promoted groups to rule out the potential influence of important baseline differences between the groups.

Matching procedure

We matched one retained child with one promoted child based on their propensity scores using SAS 8.0 PROC ASSIGN (Ming & Rosenbaum, 2001). PROC ASSIGN matches retained children with promoted children so that the sum of distances between the propensity scores within each of the matched pairs was minimized. To avoid any possibility of matching two children with propensity scores far away from each other, we imposed *maximum* caliper distance = .025. That is, any pair of retained and promoted children who differed in their propensity score by more than .025 could not be matched with each other. A total of 103 pairs (206 children) were successfully matched using this method. Note that there are 165 retained children in our sample, which means that the

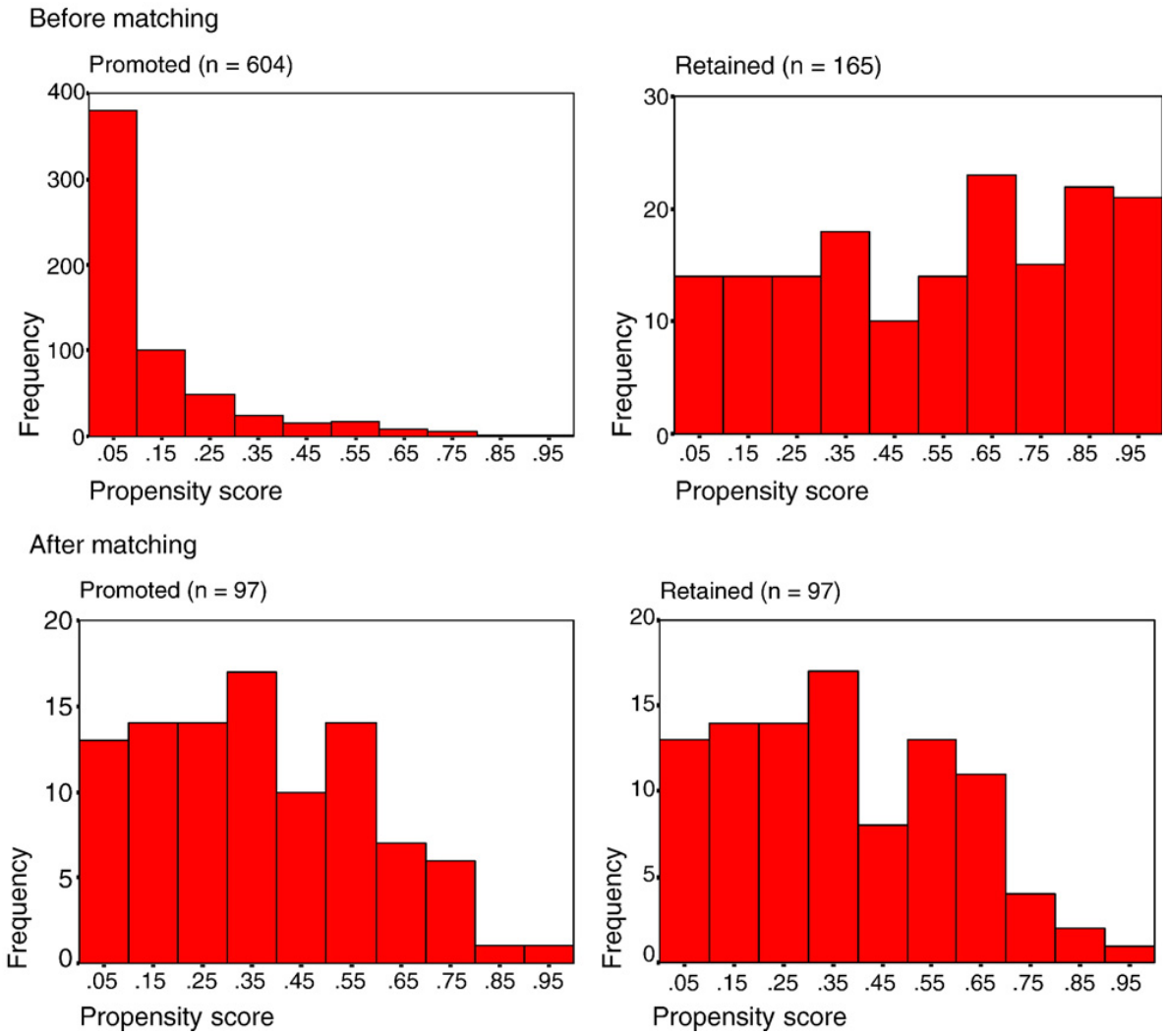

Fig. 1. The distributions of propensity score for promoted and retained groups before and after matching. Note. The scale of the *Y*-axis (Frequency) differs for the Promoted and Retained groups before matching, but is the same after matching.

maximum number of 1:1 pairs that could be matched is 165. The reason why we ended up with only 103 pairs is that we imposed very small caliper distance to achieve a very high quality of 1:1 matching on propensity scores. We could have retained more participants in our sample; however, this would have seriously compromised the quality of matches and introduced potential bias. We excluded six pairs from the growth curve analysis because at least one pair member had no data on either the WJ math or reading scores. For the remaining 97 matched pairs, the propensity score for the retained group ranged from .003 to .906 with mean = .365 and SD = .226. The propensity score for the promoted group ranged from .003 to .918 with mean = .364 and SD = .226. The mean within-pair distance in propensity score was .006 with SD = .008. Comparison of the two top panels (before matching) with the two bottom panels (after matching) of Fig. 1 shows that matching minimized the substantial pre-existing differences in propensity scores between the retained and promoted groups.

Table 1
Descriptive statistics for the time variable

	N	Minimum	Maximum	Mean	SD
Time1[1]	189	.16	1.52	.51	.27
Time2[2]	165	1.20	2.08	1.44	.19
Time3[3]	166	2.12	3.30	2.35	.20
Valid *N* (listwise)	151				

[1]Time1 = time in school (in years) for the first measurement.
[2]Time2 = time in school (in years) for the second measurement.
[3]Time3 = time in school (in years) for the third measurement.
Note. A small number of assessments were conducted as early as possible during the following school year in cases when the child could not be located (e.g., when the family moved) or a Spanish-speaking interviewer could not be matched with the child's available times.

Data analysis

Three-level linear growth curve models (Singer & Willett, 2003) were estimated separately for the WJ math and reading scores using SAS 8.0 PROC Mixed. Full information maximum likelihood estimation was used, which provides unbiased estimates with minimal standard errors when data are missing at random (Schafer & Graham, 2002). Level 1 (within-individual, Eq. (1)) captures the linear growth trajectory for each individual over three years. At level 1, the time variable TIME_{tip}, which represents each child's elapsed time in school since the beginning of the school year for grade 1 (August 24, 2001 and 2002, respectively for cohorts 1 and 2) to their exact dates of measurement on the WJ tests, was used to predict the WJ scores. The descriptive statistics for the time variable are shown in Table 1. We scaled the time variable so that the intercept (at $\mathrm{TIME}_{tip}=0$) would represent the student's predicted level of WJ math or reading on the first day of school in grade 1. Level 2 (between-individual, Eq. (2)) captures the variation in individual intercepts and slopes across individuals. At level 2, the grade retention status following first grade (coded 1 = retained, 0 = promoted) was used to predict the initial status and slope for each individual. Because matching produces dependency in the data, Level 3 (between-pair, Eq. (3)) was added to take into account the within-pair correlation for individual intercepts and slopes (clustering). This procedure adjusts for biased estimates of the standard errors caused by the dependency, thus leading to more accurate significance tests for the parameter estimates. The effect of retention on the slope (γ_{110}, the estimated annual rate of gain) is of primary interest. The equations that comprise the basic three-level model are specified below (Singer & Willett, 2003).

$$\text{Level}-1: \quad Y_{tip} = \pi_{0ip} + \pi_{1ip}\mathrm{TIME}_{tip} + e_{tip}; \qquad e_{tip} \sim N(0, \sigma^2). \tag{1}$$

$$\begin{aligned} \text{Level}-2: \quad \pi_{0ip} &= \beta_{00p} + \beta_{01p}\mathrm{RETENTION}_{ip} + r_{0ip}; \\ \pi_{1ip} &= \beta_{10p} + \beta_{11p}\mathrm{RETENTION}_{ip} + r_{1ip}; \\ \begin{bmatrix} r_{0ip} \\ r_{1ip} \end{bmatrix} &\sim \mathrm{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{\pi 00} & \tau_{\pi 01} \\ \tau_{\pi 10} & \tau_{\pi 11} \end{bmatrix} \right). \end{aligned} \tag{2}$$

$$\text{Level}-3: \quad \beta_{00p}=\gamma_{000}+u_{00p};\beta_{10p}=\gamma_{100}+u_{10p};$$
$$\beta_{01p}=\gamma_{010};\ \beta_{11p}=\gamma_{110};\begin{bmatrix}u_{00p}\\u_{10p}\end{bmatrix}\sim \text{MVN}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\tau_{\beta 00} & \tau_{\beta 01}\\ \tau_{\beta 10} & \tau_{\beta 11}\end{bmatrix}\right). \quad (3)$$

Here the subscript t indicates time point (observation 1, 2, or 3), i indicates individual, p indicates pair. γ_{000} and γ_{100} represent the grand mean intercept and slope for the promoted group. γ_{010} and γ_{110} represent the estimated effect of grade retention on the intercept and slope. e_{tip} represents level-1 residuals, which were assumed to follow a normal distribution with mean = 0 and homogeneous variance across three years. r_{0ip} and r_{1ip} represent level-2 residuals in individual intercepts and slopes. u_{00p} and u_{10p} represent level-3 deviations of pair mean intercepts and slopes from the grand mean intercept and slope for promoted children. Both level-2 residuals and level-3 deviations are assumed to follow a multivariate normal distribution with 0 means and unstructured covariance structure.

Next we explored the nine potential moderators described before for WJ math and reading score. Table 2 shows the descriptive statistics for 8 continuous moderators (only LEP is a binary variable). These potential moderator variables were added into level-2 of the growth model one at a time as shown in Eq. (4). We centered the 8 continuous variables at their grand means. The moderating effects of these variables were captured by the coefficients associated with their interaction with grade retention (γ_{130} for the slope).

$$\begin{aligned}\text{Level}-2: \quad \pi_{0ip} &= \beta_{00p}+\beta_{01p}\text{RETENTION}_{ip}+\beta_{02p}\text{MODERATOR}_{ip}\\ &\quad +\beta_{03p}\text{MODERATOR}_{ip}\times\text{RETENTION}_{ip}+r_{0ip};\\ \pi_{1ip} &= \beta_{10p}+\beta_{11p}\text{RETENTION}_{ip}+\beta_{12p}\text{MODERATOR}_{ip}\\ &\quad +\beta_{13p}\text{MODERATOR}_{ip}\times\text{RETENTION}_{ip}+r_{1ip}. \quad (4)\end{aligned}$$

$$\begin{aligned}\text{Level}-3: \quad &\beta_{00p}=\gamma_{000}+u_{00p};\ \beta_{10p}=\gamma_{100}+u_{10p};\\ &\beta_{01p}=\gamma_{010};\ \beta_{02p}=\gamma_{020};\ \beta_{03p}=\gamma_{030};\ \beta_{11p}=\gamma_{110}; \quad (5)\\ &\beta_{12p}=\gamma_{120};\ \beta_{13p}=\gamma_{130}.\end{aligned}$$

Here γ_{000} and γ_{100} represent the estimated mean intercept and slope for promoted group with a value of 0 (=mean) on the moderator. γ_{010} and γ_{110} represent the average (main) effects of grade retention on the intercept and slope. γ_{020} and γ_{120} represent the

Table 2
Descriptive statistics for the continuous potential moderators

Moderators	Promoted		Retained	
	Mean	SD	Mean	SD
Age at eligibility determination	7.58	.36	7.50	.29
Parent rated conduct problems	1.79	2.09	1.56	2.28
Parent rated hyperactivity	4.16	2.99	2.79	2.90
Teacher rated conduct problems	1.66	2.25	1.11	1.88
Teacher rated hyperactivity	3.70	3.13	2.62	2.89
Teacher–student relationship	3.91	.83	3.99	.76
Resilient personality	−.28	.88	−.12	.83
Peer rated externalizing problems	.25	.98	.29	.99

Table 3
The parameter estimates from the three-level model without moderators for WJ math and reading score

	Estimate	S.E.	Wald z	p
WJ math score				
Mean intercept for promoted (γ_{000})	454.51*	1.48	307.99	<.001
Effect of retention on the intercept (γ_{010})	.19	2.08	.09	.93
Mean slope for promoted (γ_{100})	13.24*	.59	22.53	<.001
Effect of retention on the slope (γ_{110})	−2.88*	.87	−3.30	.001
WJ reading score				
Mean intercept for promoted (γ_{000})	413.77*	2.35	179.79	<.001
Effect of retention on the intercept (γ_{010})	−4.05	3.26	−1.24	.22
Mean slope for promoted (γ_{100})	26.44*	1.09	24.19	<.001
Effect of retention on the slope (γ_{110})	−2.02	1.63	−1.24	.22

*p<.05.

average (main) effects of moderator on the intercept and slope. γ_{030} and γ_{130} represent the interaction effects of the moderator and retention status on the intercept and slope. The interpretations and covariance structures of level-1 and level-2 residuals are similar to the

Table 4
Moderating effects of 9 potential moderators (γ_{130}), average (main) effects of retention (γ_{110}) and moderators (γ_{120}) for WJ math and reading score

Moderator	WJ math score			WJ reading score		
	γ_{110}[1]	γ_{120}[2]	γ_{130}[3]	γ_{110}[1]	γ_{120}[2]	γ_{130}[3]
Age at eligibility determination	−2.97*	−1.42*	−2.92	−2.39	−7.16*	−2.13
	(.88)	(.71)	(2.70)	(1.58)	(2.88)	(4.89)
Limited English proficiency	−1.87*	3.79*	−6.92*	−.32	7.17*	−11.44*
	(.93)	(1.62)	(2.53)	(1.73)	(2.98)	(4.72)
Parent rated conduct problems	−2.09*	−.38	.56	−2.17	−1.14*	2.00*
	(.87)	(.28)	(.40)	(1.63)	(.54)	(.78)
Parent rated hyperactivity	−2.22*	−.04	.26	−2.02	.34	.32
	(.30)	(.20)	(.30)	(1.88)	(.59)	(.59)
Teacher rated conduct problems	−3.05*	−.14	.37	−3.06	.94	.99
	(.91)	(.26)	(.42)	(1.68)	(.76)	(.76)
Teacher rated hyperactivity	−3.09*	−.07	.26	−3.04	−.25	.93
	(.90)	(.24)	(.30)	(1.69)	(.34)	(.55)
Teacher–student relationship	−3.19*	.96	−.81	−2.74	.80	−3.66
	(.90)	(.74)	(1.17)	(1.67)	(1.33)	(2.10)
Resilient personality	−3.05*	.40	−.22	−3.07	.84	−1.75
	(.88)	(.67)	(1.08)	(1.67)	(1.22)	(1.98)
Peer rated externalizing problems	−2.67*	−.66	2.59*	−1.46	−.01	2.72
	(1.09)	(1.60)	(1.09)	(1.82)	(1.19)	(1.80)

*p<.05; S.E.s are in parentheses.
[1]γ_{110} denotes the average (main) effect of grade retention on the slope controlling for each moderator.
[2]γ_{120} denotes the average (main) effect of moderator on the slope controlling for grade retention.
[3]γ_{130} denotes moderating effect of each potential moderator (or the interaction effect between grade retention and the moderator).

previous model. u_{00p} and u_{10p} represent level-3 deviations of pair mean intercepts and slopes from the grand mean intercept and slope for promoted children with a value of 0 (=mean) on the moderator.

Results

The results obtained from the basic growth model (without moderators) for WJ math and reading scores are summarized in Table 3. The moderating effects of the 9 potential moderators (γ_{130}) from the growth model in Eqs. (4) and (5) are summarized in Table 4 (the 4th and 7th columns). Also, to see whether the effects of grade retention on the slope will change as moderators were added compared to the overall effects of grade retention in Table 3, we also reported the average (main) effects of grade retention and the moderators on the slope in Table 4.

WJ math score

As shown in Table 3, grade retention had no effect on the intercept (Wald $z=.09$, ns), indicating that the initial status on the WJ Math test did not differ between the retained and promoted groups. Matching on propensity scores achieved initial equivalence of the promoted and retained groups on WJ math scores. Of importance, Grade Retention had a

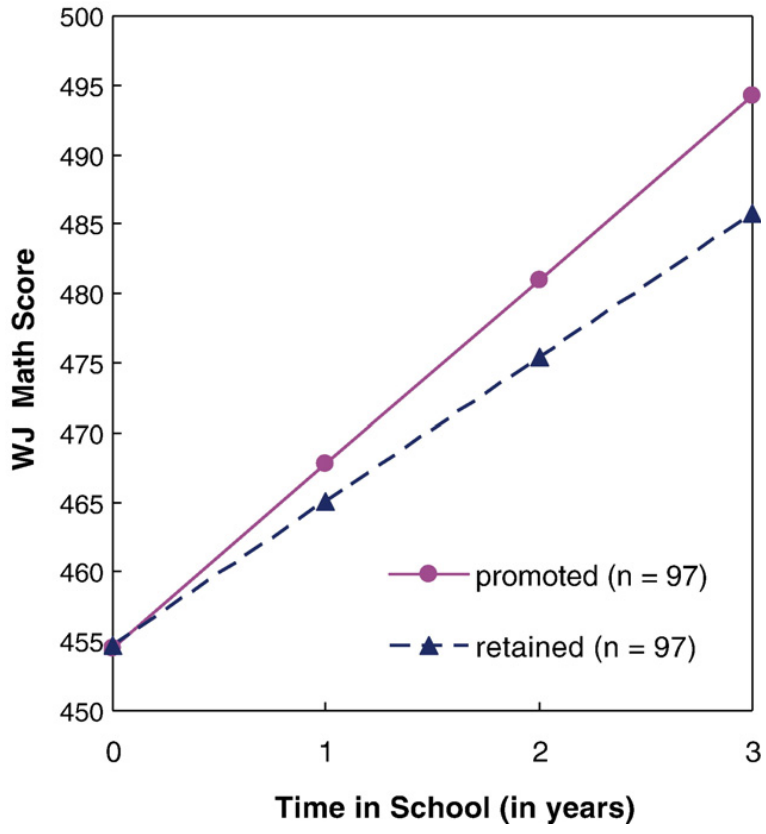

Fig. 2. Estimated linear growth curves of WJ math score for retained and promoted children.

negative effect on the slope ($\gamma_{110}=-2.88$, Wald $z=-3.30$, $p=.001$), indicating that on average the retained children had a *lower* rate of gain of 2.88 points per year on the WJ Math test than the promoted children. A gain of 2.88 points on the WJ math represents 31% of the average annual increase for children ages 8–9. The estimated annual rate of gain on WJ math score was 13.24 (γ_{100}, Wald $z=21.36$, $p<.001$) for promoted children, while it was 10.36 ($\gamma_{100}+\gamma_{110}$) for retained children. These results are illustrated in Fig. 2, which portrays the estimated growth lines for the overall retained and promoted groups.

Among the nine potential moderator variables we examined, two variables had a significant interaction with grade retention in predicting the slope for WJ math score, as shown in Table 4. The interaction between LEP status and grade retention had a negative effect on the slope ($\gamma_{130}=-6.92$, Wald $z=-2.74$, $p=.006$), indicating that retention showed a more negative effect on the slope for LEP than for non-LEP children (see Fig. 3). The interaction between peer rated externalizing problems and grade retention had a positive effect on the slope ($\gamma_{130}=2.59$, Wald $z=2.39$, $p=.02$). As shown in Fig. 4, the negative impact of retention was stronger for children with low levels of externalizing problems than for children with higher levels of externalizing problems.

As shown in Table 4, the main effects of grade retention on the slope for WJ math score were negative and statistically significant after controlling for each of the 9 moderators. These results are consistent with the overall effect of grade retention on the slope for WJ math score shown in Table 3.

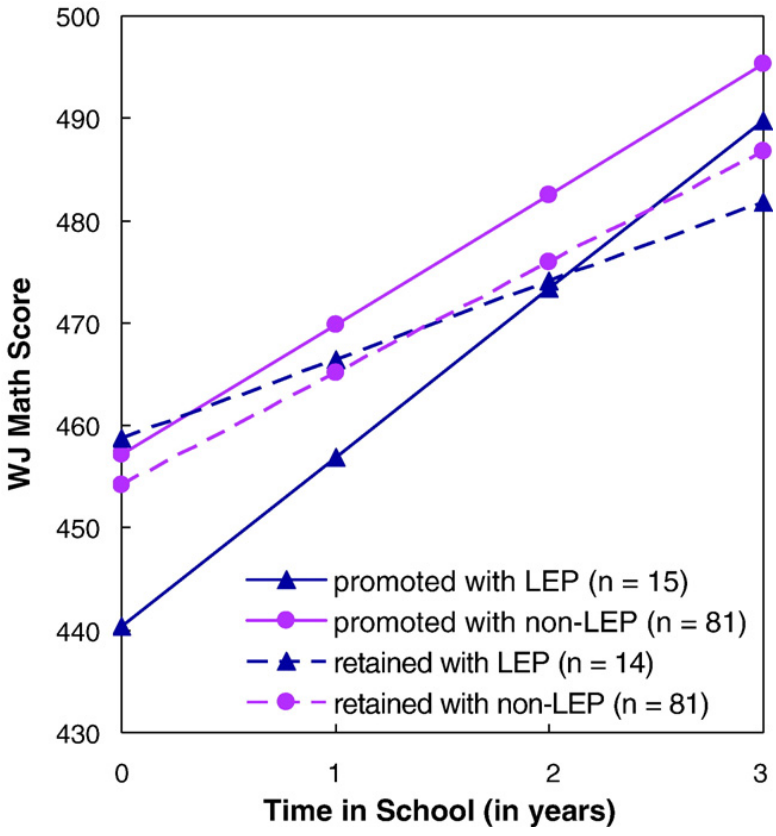

Fig. 3. Estimated linear growth curves of WJ math score for retained and promoted children with limited and adequate English proficiency.

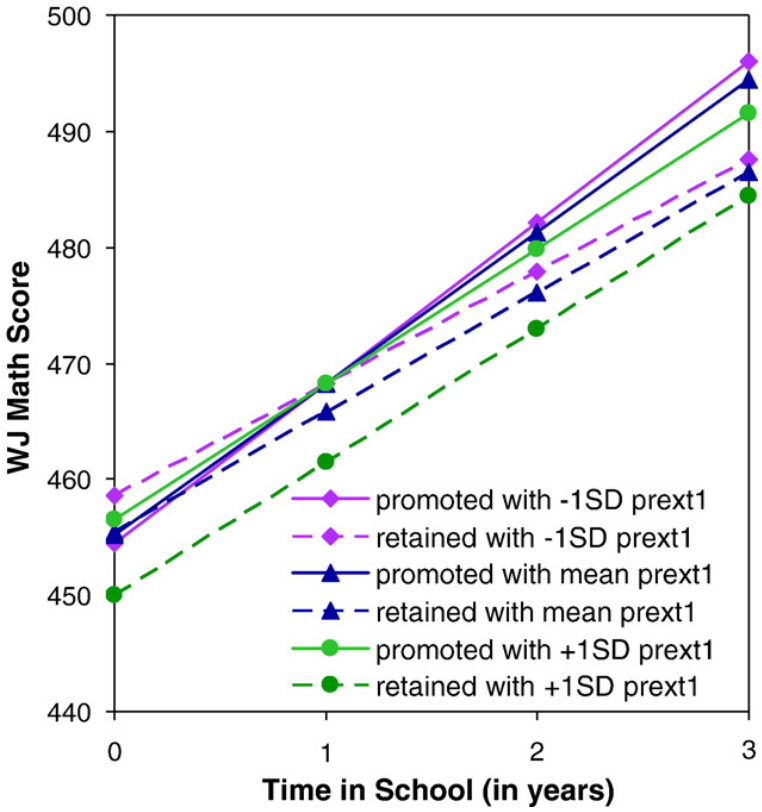

Fig. 4. Estimated linear growth curves of WJ math score for retained and promoted children with −1SD, mean and +1SD of peer rated externalizing problems (prext1).

WJ reading score

As shown in Table 3, grade retention had no effect on the intercept (Wald $z=-1.24$, ns), indicating that initial status on the WJ reading test also did not differ between the retained and promoted groups. This provides further evidence of the success of the propensity score matching procedure. Although grade retention also tended to have a negative effect on the slope ($\gamma_{110}=-2.02$), the effect was not statistically significant (Wald $z=-1.24$, $p=.22$).

Among the 9 moderators we examined, we identified two variables that had a significant interaction with grade retention in predicting the slope for WJ reading score (see Table 4). The interaction between LEP status and grade retention had a negative effect on the slope ($\gamma_{130}=-11.44$, Wald $z=-2.42$, $p=.02$), indicating that grade retention showed a more negative effect on the slope for LEP children than for non-LEP children (see Fig. 5). The interaction between parent-rated conduct problems and grade retention had a positive effect on the slope for WJ reading score ($\gamma_{130}=2.00$, Wald $z=2.58$, $p=.01$), indicating that grade retention showed less negative effect on the slope for those children with higher initial parent ratings on conduct problems. The estimated growth curves associated with this moderator shows a very similar pattern as those associated with peer-rated externalizing problems in Fig. 4 for math scores.

As shown in Table 4, the main effects of grade retention on the slope for WJ reading score did not differ significantly from 0 after controlling for each of the 9 moderators. These

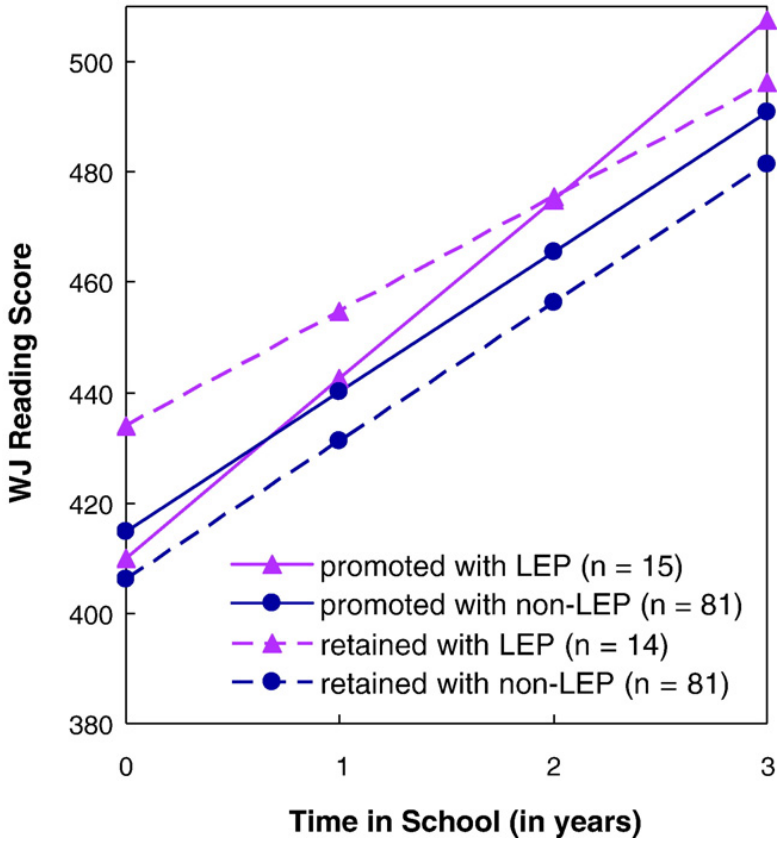

Fig. 5. Estimated linear growth curves of WJ reading score for retained and promoted children with limited and adequate English proficiency.

results are also consistent with the overall effect of grade retention on the slope for WJ reading score shown in Table 3.

Discussion

The results showed that grade retention decreased the rate of growth for mathematical skills but did not have a significant effect on reading skills in the initial two years following retention. This negative effect on the rate of growth for mathematical skills still holds after controlling for other variables that also had influence on the rate of growth for mathematical skills (e.g. age, LEP, see Table 4). The overall negative effect of retention on math but not on reading is most likely due to the fact that Texas's curricular guidelines for math instruction indicate greater grade differentiation in new concepts taught in math than in reading in grades 1–3 (Texas Education Agency, 2006b). Thus, retained students' lower growth is likely due to lack of instruction in new concepts.

These results provided no evidence of overall beneficial short-term effects of grade retention. Three aspects of this finding are of particular importance. First, this failure to find positive effects occurred in young elementary school children who were retained in first grade. At this age, many of the hypothesized mechanisms that may contribute to negative effects of retention (e.g., being labeled as being "held back" by peers) should be minimal. Second, results were found on the WJ broad math and reading scores, well-validated

measures of academic competence. Because our assessment was independent of state and school district testing, the teachers could not "teach to the test" to improve scores without improving general competence. Third and of most importance, this study used retained and promoted children from an initial at risk sample who were then closely matched on their propensity scores. The propensity score is a device for constructing matched sets when we need to balance two groups on many covariates (Rosenbaum, 2002; West & Thoemmes, in press). In our case, matching on propensity score provides a strong control for potential baseline differences on 72 measured background characteristics between retained and promoted children. These background characteristics were based on a comprehensive assessment of demographic, academic ability, and psychosocial adjustment variables known to be related to grade retention. Although we cannot definitively rule out the possibility that pre-existing differences on some unmeasured variable potentially accounts for the results, this possibility is greatly reduced given the comprehensiveness of the baseline assessment and the balance achieved on the covariates (West & Thoemmes, in press). The use of propensity scores substantially improved the internal validity of the effects detected in this study.

We also identified three moderators of the effect of retention: LEP, peer rated externalizing problems and parent rated conduct problems. For both WJ math and WJ reading, the negative effect of grade retention on the growth rate was strengthened for LEP children. Several possible explanations for this effect might be investigated in future studies. For example, 16 of the 29 LEP children were in bilingual classes. There is typically only one bilingual classroom per grade in a school; thus, more than one-third of the LEP children who were retained completed their second year in first grade with the same teacher. None of the promoted LEP children had the same teacher the second year. Perhaps repeating first grade with the same teacher has a dampening effect on growth in reading due to a reduction in the extent to which children are exposed to different teaching methods and curricula. Repeating the grade with the same teacher may also increase the probability of replicating a less than optimal teacher–student match. We caution that these results were obtained based on a small sample of LEP children so that replication is needed before reaching firm conclusions.

With respect to peer-rated conduct problems, as expected, the positive effect of grade promotion was strengthened as children's conduct problems decreased. Growth in math was steeper for promoted children with low peer-rated externalizing problems, relative to retained children with low peer-rated externalizing problems. Growth in reading was steeper for promoted children with low parent-rated externalizing problems, relative to retained children with low parent-rated externalizing problems. Taken together, these findings suggest that children who have better behavioral regulation benefit more from being promoted, perhaps because they are better able to cope with the more challenging environment of the second grade. That is, their good behavioral regulation allows them to "rise to the occasion."

As for the other six potential moderators, although they did not show significant moderating effects in our study, the directions of those effects were all consistent with our hypotheses (see Table 4). As we expected, retention seems to be less beneficial for older children than for younger children. Similar to parent and peer rated externalizing problems, teacher rated externalizing problems tends to reduce the positive effect of promotion. In

addition, the positive effect of promotion seems to be strengthened as teacher–student relationship quality and resilient personality increased. These results provide us with clues as to which groups of children are more or less vulnerable to the grade retention with respect to their math and reading skills. We fully acknowledge that these results are clearly exploratory and require replication before strong conclusions can be reached.

Finally, our study has two primary limitations. In the moderator analyses, the sample size of the LEP group was small. In future work on grade retention, LEP students should be oversampled to provide a better understanding of the effects of retention on this increasingly important group of students. In addition, the present analyses could only investigate the short-term linear effects of grade retention given three waves of data collection. It is possible that the longer term effects of grade retention may take a different form. As part of an ongoing longitudinal study, our future work will permit us to investigate more complex, longer term effects of grade retention on WJ math and reading scores. The collection of additional measurement waves will permit us to estimate nonlinear trajectories and piecewise trajectories in which retention has an initial short term effect, followed by a different longer term effect.

References

Alexander, K. L., Entwisle, D. R., & Dauber, S. L. (1993). First-grade classroom behavior: Its short- and long-term consequences for school performance. *Child Development*, *64*, 801−814.

Alexander, K. L., Entwistle, D. R., & 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.

Bali, V. A., Anagnostopoulos, D., & Roberts, R. (2005). Toward a political explanation of grade retention. *Educational Evaluation and Policy Analysis*, *27*, 133−155.

Beebe-Frankenberger, M., Bocian, K. M., MacMillan, D. L., & Gresham, F. M. (2004). Sorting second-grade students: Differentiating those retained from those promoted. *Journal of Educational Psychology*, *96*, 204−215.

Bracken, B. A., & McCallum, R. S. (1998). *Universal nonverbal intelligence test: Examiner's manual.* Itasca, IL: Riverside.

Buhrmester, D., & Furman, W. (1987). The development of companionship and intimacy. *Child Development*, *54*, 1386−1399.

Caspi, A., Block, J., Block, J. H., & Klopp, B. (1992). A "common-language" version of the California Child Q-set for personality assessment. *Psychological Assessment*, *4*, 512−523.

Cicchetti, D., & Posner, M. I. (2005). Cognitive and affective neuroscience and developmental psychopathology. *Development and Psychopathology*, *17*, 569−575.

Cillessen, A. H. N., & Bukowski, W. M. (2000). *Recent advances in the measurement of acceptance and rejection in the peer system.* San Francisco, CA: Jossey-Bass.

Dauber, S. L., Alexander, K. L., & Entwisle, D. R. (1993). Characteristics of retainees and early precursors of retention in grade: Who is held back? *Merrill-Palmer Quarterly*, *39*, 326−343.

Dickey, W. C., & Blumberg, S. J. (2004). Revisiting the factor structure of the Strengths and Difficulties Questionnaire: United States, 2001. *Journal of the American Academy of Child and Adolescent Psychiatry*, *43*, 1159−1167.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, *38*, 581−586.

Goodman, R. (2001). Psychometric properties of the Strengths and Difficulties Questionnaire. *Journal of Child Psychology and Psychiatry*, *40*, 791−799.

Grissom, J., & Shepard, L. (1989). Repeating and dropping out of school. In L. A. Shepard & M. L. Smith (Eds.), *Flunking grades: Research and policies on retention* (pp. 16−33). London: The Falmer Press.

Hill, C. R., & Hughes, J. N. (in press). An examination of the convergent and discriminant validity of the Strengths and Difficulties Questionnaire. *School Psychology Quarterly.*

Holmes, C. T. (1989). Grade-level retention effects: A meta-analysis of research studies. In L. A. Shepard & M. L. Smith (Eds.), *Flunking grades: Research and policies on retention* (pp. 16−33). London: The Falmer Press.

Hughes, J. (1990). Assessment of children's social competence. In C. R. Reynolds & R. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children* (pp. 423−444). New York, NY: Guilford.

Hughes, J. N., Cavell, T. A., & Willson, V. (2001). Further evidence of the developmental significance of the teacher–student relationship. *Journal of School Psychology*, *39*, 289−302.

Hughes, J. N., & Kwok, O. (2006). Classroom engagement mediates the effect of teacher–student support on elementary students' peer acceptance: A prospective analysis. *Journal of School Psychology*, *43*, 465−480.

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

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

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102−138). (2nd ed.). New York: Guilford.

Karweit, N. (1999). Retention policy. In M. Aikin (Ed.), *Encyclopedia of educational research* (pp. 1114−1118). New York: Macmillan.

Khoo, S. -T., West, S. G., Wu, W., & Kwok, O. -M. (2006). Longitudinal methods. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 301−317). Washington, DC: American Psychological Association.

Kwok, O., Hughes, J. N., & Luo, W. (2006). Role of resilient personality on lower achieving first grade students' current and future achievement. *Journal of School Psychology*, *45*, 61−82.

Lerner, R. M. (1989). Developmental contextualism and the life-span view of person–context interaction. In M. H. Bornstein & J. S. Bruner (Eds.), *Interaction in human development. Crosscurrents in contemporary psychology* (pp. 217−239). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Mantzicopoulos, P. (2003). Flunking kindergarten after head start: An inquiry into the contribution of contextual and individual variables. *Journal of Educational Psychology*, *95*, 268−278.

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

Meehan, B. T., Hughes, J. N., & Cavell, T. A. (2003). Teacher–student relationships as compensatory resources for aggressive children. *Child Development*, *74*, 1145−1157.

Meisels, S. J., & Liaw, F. R. (1993). Failure in grade: Do retained students catch up? *Journal of Educational Research*, *8*, 69−77.

Ming, K., & Rosenbaum, P. A. (2001). A note on optimal matching with variable controls using the assignment algorithm. *Journal of Computational and Graphical Statistics*, *10*, 455−463.

National Center for Education Statistics. (2006). *The condition of education 2006: Indicator 7: Language minority school-age children.*Washington, DC: Author. Accessed 12-29-06 from http://nces.ed.gov/programs/coe/2006/section1/indicator07.asp

Pagani, L., Tremblay, R. E., Vitaro, F., Boulerice, B., & McDuff, P. (2001). Effect of grade retention on academic performance and behavioral development. *Development and Psychopathology*, *13*, 297−315.

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

Pierson, L. H., & Connell, J. P. (1992). Effects of grade retention on self-system processes, school engagement, and academic performance. *Journal of Educational Psychology*, *84*, 300−307.

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

Rosenbaum, P. A. (2002). *Observational studies* (2nd Ed.). New York: Springer.

Rosenbaum, P. A., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41−55.

Sameroff, A. J. (1989). Principles of development and psychopathology. In A. Sameroff & R. Emde (Eds.), *Relationship disturbances in early childhood* (pp. 17−32). New York: Basic Books.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147−177.

Silberglitt, B., Jimerson, S. R., Burns, M. K., & Appleton, J. J. (2006). Does the timing of grade retention make a difference? Examining the effects of early versus late retention. *School Psychology Review*, *35*, 134−141.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York: Oxford.

Terry, R. (1999). *Measurement and scaling issues in sociometry: A latent trait approach*. Paper presented at the biennial meeting of the Society for Research in Child Development, Albuquerque, New Mexico.

Texas Education Agency. (2006a). *Texas Education Agency Academic Excellence Indicator System, 2006.* Retrieved June 8, 2006 from http://www.tea.state.tx.us/perfreport/aeis/2001/state.html

Texas Education Agency. (2006b). *Texas Essential Knowledge and Skills: Chapters 110 (English Language Arts and Reading) and 111 (Mathematics).* Austin, TX: Author. Accessed 12-20-06 from http://www.tea.state.tx.us/teks

Tomchin, E. M., & Impara, J. C. (1992). Unraveling teachers' beliefs about grade retention. *American Educational Research Journal*, *29*, 199–123.

West, S. G., Biesanz, J. C., & Pitts, S. C. (2000). Causal inference and generalization in field settings: Experimental and quasi-experimental designs. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in personality and social psychology* (pp. 40−84). New York: Cambridge University Press.

West, S.G., & Thoemmes, F. (in press). Equating groups. In J. Brannon, P. Alasuutari, & L. Bickman (Eds.), *Handbook of social research methods.* London: Sage.

Willson, V. L., & Hughes, J. N. (2006). Retention of Hispanic/Latino students in first grade: Child, parent, teacher, school, and peer predictors. *Journal of School Psychology*, *44*, 31−49.

Willson, V. L., Hughes, J. N., Gleason, K. A., & Wong, S. W. (submitted for publication). *Who is retained in first grade? A psychosocial perspective.* Manuscript.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock–Johnson tests of achievement: Standard and supplemental batteries.* Allen, TX: DLM Teaching Resources.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *WJ-III Tests of Achievement.* Itasca, IL: Riverside Publishing.

Woodcock, R. W., & Muñoz-Sandoval, A. F. (1993). *Woodcock–Muñoz Language Survey.* Chicago: Riverside.

Woodcock, R. W., & Muñoz-Sandoval, A. F. (1996). *Batería Woodcock–Muñoz Pruebas de Aprovechamiento—Revisada.* Itasca, IL: Riverside.

Woodock, R. W., & Muñoz-Sandoval, A. F. (2001). *Comprehensive Manual. Woodcock–Muñoz Language Survey Normative Update.* Itasca, IL: Riverside Publishing.