Journal of Educational Psychology
2010, Vol. 102, No. 1, 135–152

0022-0663/10/$12.00 DOI: 10.1037/a0016664

Effect of Grade Retention in First Grade on Psychosocial Outcomes

Wei Wu
University of Kansas

Stephen G. West
Arizona State University

Jan N. Hughes
Texas A&M University

In a 4-year longitudinal study, the authors investigated effects of retention in first grade on children's externalizing and internalizing behaviors; social acceptance; and behavioral, cognitive, and affective engagement. From a large multiethnic sample ($n = 784$) of children below the median on literacy at school entrance, 124 retained children were matched with 251 promoted children on the basis of propensity scores (probability of being retained in first grade estimated from 72 baseline variables). Relative to promoted children, retained children were found to benefit from retention in both the short and longer terms with respect to decreased teacher-rated hyperactivity, decreased peer-rated sadness and withdrawal, and increased teacher-rated behavioral engagement. Retained children had a short-term increase in mean peer-rated liking and school belongingness relative to promoted children, but this advantage showed a substantial decrease in the longer term. Retention had a positive short-term effect on children's perceived school belonging and a positive longer term effect on perceived academic self-efficacy. Retention may bestow advantages in the short-term, but longer term detrimental effects on social acceptance may lead to the documented longer term negative effects of retention.

Keywords: grade retention, growth curve model, propensity score, optimal matching, psychosocial outcomes

For at least the past 4 decades, grade retention, the practice of having a student who has been in a given grade level for a full school year to remain at that level for a subsequent school year, has been a common but controversial educational practice (Holmes, 1989; Jackson, 1975; Lorence, 2006). With the increase in high stakes testing that began in the early 1990s and the related requirement that students meet grade-level academic competencies, grade retention has increased in frequency (Bali, Anagnostopoulos, & Roberts, 2005; Gootman, 2005; Roderick & Nagaoka, 2005). In 2004, U.S. Census data revealed that 9.6% of U.S. youths ages 16–19 years had been retained in grade one or more times. In Texas, the location of the current study, during the 2003–2004 school year, retention in first grade, the focal grade in the current study, was 6.4%, compared with 5.8% in 1994–1995, prior to implementation of high stakes testing (Texas Education Agency, 2005). These data suggest an increase of more than 10%, as compared with a decade earlier, in the proportion of children being retained in first grade.

Given the prevalence of grade retention, one might expect that it enjoys strong empirical support. On the contrary, the current empirical evidence of the effects of grade retention on both academic and social–emotional adjustment is inconsistent but is widely characterized in published literature as negative (for meta-analyses, see Holmes, 1989; Jimerson, 2001a; for narrative reviews, see Jimerson, 2001b; Shepard, Smith, & Marion, 1996). However, the widely acknowledged methodological limitations of published studies make it difficult to reach firm conclusions regarding a causal role for grade retention on achievement (Hong & Raudenbush, 2005; Lorence, 2006; Wu, West, & Hughes, 2008b). The most challenging design issue in studies of the effect of grade retention is that of making causal inferences about grade retention in the absence of a randomized experimental design (Campbell & Stanley, 1966; Shadish, Cook, & Campbell, 2002; West & Thoemmes, 2008). 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 postretention outcomes. Retention researchers have attempted to minimize the influence of such confounding variables that may be related to both the retention decision and the academic and social outcomes. Commonly, researchers have addressed this issue by selecting promoted students who are matched to retained students on achievement-relevant variables or by statistically controlling for preretention variables that predict achievement, such as family variables, child characteristics, or academic performance. However, these attempts often fall short. In a qualitative review of 18

Wei Wu, Department of Psychology, University of Kansas; Stephen G. West, Department of Psychology, Arizona State University; Jan N. Hughes, Department of Educational Psychology, Texas A&M University.

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

Correspondence concerning this article should be addressed to Wei Wu, Department of Psychology, University of Kansas, Lawrence, KS 66045-7556, or Jan N. Hughes, Department of Educational Psychology, Texas A&M University, College Station, TX 77843-4225. E-mail: wwei@ku.edu or jhughes@tamu.edu

studies included in Jimerson's (2001a) often-cited meta-analysis, Lorence (2006) judged that only four studies had both adequate comparison groups and statistical controls.

In addition to methodological limitations, research on the effects of grade retention is limited by its largely atheoretical basis. Virtually absent in the literature are prospective studies that address developmentally informed hypotheses regarding how grade retention affects achievement. Researchers have speculated that the effects of grade retention on achievement and achievement-related outcomes (e.g., school completion) are mediated by more proximal psychosocial variables (Alexander, Entwisle, & Dauber, 2003; Pagani, Tremblay, Vitaro, Boulerice, & McDuff, 2001). Consistent with developmental systems and transactional theories of development (Lerner, 1998; Sameroff, 1975), it is probable that retention in the elementary grades influences multiple risk and protective processes. These effects may unfold over time, affecting different dimensions of functioning during different developmental periods, with some effects not emerging until later developmental periods (Coie et al., 1993).

Consistent with such dynamic theories of development, in one of the most detailed and longest studies of the effects of retention in first grade on academic and behavioral outcomes through high school, Alexander, Entwisle, and Dauber (2003) found that early-retained children in Baltimore schools, relative to matched promoted children, were more likely to drop out of school in adolescence, despite performing better in their coursework than promoted children. Alexander et al. (2003) concluded: "Retention, so far as we can determine, does not impede . . . children academically or assault their self-esteem in the early years, yet something about the experience apparently weakened repeaters' attachment to school" (p. ix). Identifying the effects of early retention on students' psychosocial functioning and social relationships at school during the elementary grades is a necessary step in building more comprehensive developmental models that test theoretically and empirically informed hypotheses regarding the processes responsible for retention's effects on both psychosocial and academic outcomes in later grades.

In the present study, we investigated the effects of retention in first grade on children's behavior problems, social acceptance, and engagement. We carefully matched retained and nonretained children using modern propensity score procedures (Rosenbaum, 2002) that closely equate nonrandomized groups. Our study was informed by the prior literature, a transactional perspective, and theorizing based on social comparison theory (Festinger, 1954) and the big fish, little pond effect (Marsh & Craven, 2002). Children were assessed each year for 4 years, permitting us to separate the short- and longer term effects of retention, which have often differed in previous research (Alexander et al., 2003; Hong & Yu, 2008; Pierson & Connell, 1992). Through the use of these techniques, we sought to address many of the limitations of prior research and provide a clearer picture of the effects of early grade retention.

Social Comparison Theory and Retention Effects

Social comparison theory (Festinger, 1954) offers a potentially useful theoretical framework for understanding retention's short-term and longer term effects. The theory posits that individuals use cues in their environment to make inferences about their own and others' relative abilities. Classrooms provide multiple cues of children's relative abilities, including teachers' differential responses to students' responses, public praise and criticism, length of time provided for children to answer a question, graded assignments, and student grouping based on ability (Brophy, 1983; Jussim, 1986; Mac Iver, 1988; Weinstein, Marshall, Sharp, & Botkin, 1987). Children as young as first graders use these cues in making inferences about their own and peers' abilities (Stipek, 1981; Stipek & Tannatt, 1984). Social comparison theory has been used to explain the big fish, little pond phenomenon (Marsh & Craven, 2002), which refers to the boost in self-perceived competence that accompanies a downward shift in the ability level of one's social frame of reference.

Grade retention brings about such a change in reference. During the repeat year, the retained child is, on average, 1 year older than his or her grade mates and has 1 additional year of experience with the curriculum and classroom routines. Research on the short-term effects of grade retention has found that retained children improve in achievement relative to their younger grade mates (Anderson, Jimerson, & Whipple, 2005; Gleason, Kwok, & Hughes, 2007; Pierson & Connell, 1992); however, as the number of years postretention increases, these benefits diminish (Hong & Yu, 2008; Pierson & Connell, 1992; Wu, West, & Hughes, 2008a). The retained child is also likely to improve more than promoted students in the short term in meeting classroom expectations for behavioral and social competencies. Social immaturity (e.g., difficulty paying attention, excessive motor activity, poor social skills, poor emotional regulation) is a primary reason teachers give for retaining a child (Tomchin & Impara, 1992). During the early elementary grades, the self-regulatory skills important to behavioral and emotional functioning are developing rapidly (Blair, 2002; Kochanska, Murray, & Coy, 1997). With an extra year of physical maturation, the retained child's abilities to inhibit impulsive actions, to focus attention, and to conform to classroom rules are expected to improve relative to the prior year and relative to his or her (younger) classmates. In the longer term, however, the benefit of an extra year of maturation may wane.

Previous Research on Retention Effects on Psychosocial Outcomes

Empirical Studies of Retention Effects on Psychosocial Adjustment

Jimerson (2001a) conducted a meta-analysis of 16 studies published between 1990 and 1999 reporting a total of 77 effect sizes for socioemotional and behavioral adjustment. These studies yielded a mean weighted effect size of $-.22$ (Cohen's d; Cohen, 1988), indicating that promoted students fared better than did retained students. However, the large majority (86%) of the analyses yielded no statistically significant differences between retained and comparison students on socioemotional adjustment. Because nearly all of these studies failed to include adequate controls for potential baseline differences on psychosocial variables between the retained and promoted children, the results are not conclusive.

Given our ultimate interest in identifying the shorter term effects of retention on adjustment that may mediate its longer term effect on achievement, we focused our literature review on psychosocial variables for which there is empirical evidence of a link to achievement. These variables can be categorized into three psychosocial adjustment domains: behavioral adjustment (externalizing and internalizing behaviors), social acceptance, and engagement in learning (including behavioral, affective, and cognitive engagement). For each domain, we summarize evidence related to two links: (a) the potential role of the psychosocial variable in achievement and (b) the effects of retention on functioning in that psychosocial domain. We focus our review on studies published since 1990, which represent more contemporary conditions of schooling and populations of schoolchildren.

Externalizing Problems

Linkages with achievement. The relation between externalizing and achievement is complex. A number of prospective investigations have documented that high levels of externalizing problems predict lower academic performance, after controlling for previous levels of academic performance (Hinshaw, 1992; Masten et al., 2005; Miles & Stipek, 2006; Risi, Gerhardstein, & Kistner, 2003; Trzesniewski, Moffitt, Caspi, Taylor, & Maughan, 2006).

Effect of retention on externalizing behaviors. Some authors posit that the frustration and humiliation associated with repeating the curriculum, combined with one's physical size, may result in an increase in aggression and oppositional behavior (Pagani et al., 2001). Studies investigating effects of grade retention on externalizing problems have produced inconsistent findings, with the majority of studies finding no effects (Alexander, Entwisle, & Dauber, 1994; Jimerson, Carlson, Rotert, Egeland, & Sroufe, 1997; Mantzicopoulos & Morrison, 1992; McCombs Thomas et al., 1992; McCoy & Reynolds, 1999). Some researchers have reported that retention increases students' externalizing problems (Pianta, Tietbohl, & Bennett, 1997; Pagani et al., 2001), whereas at least one study documented that retention results in a decrease in externalizing problems (Gottfredson, Fink, & Graham, 1994). Variations in the timing of retention, the length of the follow-up, and the adequacy of controls for selection factors may contribute to inconsistent findings.

Internalizing Behaviors

Linkages with achievement. Poor academic performance is linked with symptoms of both anxiety (Beidel, 1991; Hembree, 1988) and depression (Herman, Lambert, Reinke, & Ialongo, 2008; Schwartz, Gorman, Duong, & Nakamoto, 2008). The association between academic performance and anxiety/depression may be due in part to reciprocal causal processes (Cole, Martin, & Powers, 1997). Anxious thoughts and feelings may disrupt attention and concentration and lead to the pursuit of performance-avoidance goals, lowering performance on academic tasks (McDonald, 2001). Anxiety related to one's academic performance is thought to lead to the pursuit of performance-avoidance goals that impede learning (Elliot, Sheldon, & Church, 1997). Depressed individuals may also avoid academic challenge and fail to persist in the face of failure, thereby failing to learn skills necessary for successful academic performance (Strauss, Lahey, & Jacobsen, 1982). Recent longitudinal studies of elementary students find that learning problems, especially reading difficulties, have a stronger effect on internalizing symptoms (anxiety and depression) than internalizing symptoms have on achievement (Ackerman, Izard, Kobak, Brown, & Smith, 2007; Maughan, Rowe, Loeber, & Stouthamer-Loeber, 2003).

Effect of retention on internalizing behaviors. Research on the effects of grade retention on internalizing symptoms of depression, anxiety, and withdrawal has produced mixed findings. Studies that did not control for preretention levels of internalizing symptoms have reported negative effects of retention on internalizing symptoms (Jimerson et al., 1997; McCombs Thomas et al., 1992; Meisels & Liaw, 1993; Pianta et al., 1997). Hong and Yu (2008) used propensity strata to adjust for selection effects and HLM to address school effects in a large data set representative of the U.S. population. Hong and Yu found that at a 2-year follow-up, the repeat-year students who had been retained in kindergarten had fewer child-reported internalizing problem behaviors than would be expected if they had been promoted. In contrast, Pagani et al. (2001) found a negative effect of retention on children's internalizing problems. Differences may be due to differences in the grade retained (kindergarten in Hong & Yu, 2008, and elementary grades in Pagani et al., 2001) or to differences in the educational systems in the United States and Quebec.

Social Acceptance

Linkages with academic achievement. Children who are more accepted and less rejected by their classmates are likely to perform better academically (Buhs & Ladd, 2001; Furrer & Skinner, 2003; Ladd, Birch, & Buhs, 1999; Zettergren, 2003). Longitudinal studies find that the positive relation between peer acceptance (or rejection) and academic achievement holds even when the effect of prior levels of achievement is statistically controlled (Buhs, 2005; Buhs, Ladd, & Herald, 2006). The effect of peer acceptance on achievement may be mediated by the effect of peer acceptance on a student's engagement in the classroom (Ladd et al., 1999) and academic self-efficacy beliefs (Flook, Repetti, & Ullman, 2005).

Effect of retention on social acceptance. The few studies on the effects of retention on peer acceptance have relied on teacher or parent reports or on the child's own perceptions of peer support rather than on peer assessments of liking for the child. The results have been inconsistent. For example, when kindergarten teachers were asked to rate classroom behavior and peer acceptance, they rated retained students as being less well liked by classmates than low-achieving promoted students (Pianta et al., 1997). In the methodologically rigorous Hong and Yu (2008) study, described above, no effect for kindergarten retention on teacher-rated peer acceptance was found 2 and 4 years postretention. In studies asking students to report on their perceived social acceptance, no differences are found between retained and ability-matched promoted students (Gottfredson et al., 1994; Hagborg, Masella, Palladino, & Shepardson, 1991; Pierson & Connell, 1992).

Engagement

Linkages with achievement. The construct of engagement in learning is widely viewed as multidimensional, encompassing behavioral (e.g., persisting on tasks, following classroom rules), affective (e.g., liking school, sense of school belonging), and cognitive (e.g., believing that one is academically capable, possessing a learning or mastery orientation) dimensions (Alexander et al., 1993; Appleton, Christenson, Kim, & Reschly, 2006; Finn, 1989; Fredericks, Blumenfeld, & Paris, 2004). Appleton et al. described the affective and cognitive dimensions of engagement as motivational processes that drive "the direction, intensity, and quality of one's energies" and behavior engagement as "energy in action" (p. 428).

Prospective studies with elementary students have documented effects of behavioral engagement on achievement, above levels of prior achievement (Alexander et al., 1993; Greenwood, 1991; Hughes, Luo, Kwok, & Loyd, 2008; Ladd et al., 1999; Skinner, Zimmer-Gembeck, & Connell, 1998). With respect to affective engagement, Valeski and Stipek (2001) found that first graders' perceived liking for school was positively associated with academic skills assessments. Ladd, Buhs, and Seid (2000) found that school liking in kindergarten fosters classroom engagement, which leads to higher achievement. An extensive body of research with older elementary and middle-school students documents longitudinal associations between students' academic self-efficacy beliefs, an aspect of cognitive engagement, and achievement (for reviews, see Pajares, 1996; Schunk & Zimmerman, 2006). Furthermore, children as young as first grade who report higher perceptions of academic competence are more behaviorally engaged in school (e.g., Hughes & Zhang, 2007).

Effect of grade retention on engagement. Findings on the effect of retention on different dimensions of engagement are mixed. A finding of no effect is most common for behavioral engagement (Ferguson, 1991; Gottfredson et al., 1994; Pierson & Connell, 1992), liking for or interest in school (Alexander et al., 1994; Hagborg et al., 1991), and academic self-efficacy (Alexander et al. 1994; McCoy & Reynolds, 1999; Phelps, Dowdell, Rizzo, Ehrlich, & Wilczenski, 1992). The few findings of negative effects tend to come from studies with poor controls for preselection differences (Hagborg et al., 1991; Pianta et al., 1997). Importantly, a study using strong controls for preretention vulnerabilities reported positive effects of retention in the kindergarten grades on academic self-efficacy 2 and 4 years later (Hong & Yu, 2008). In a longitudinal study of a predominantly African American, urban sample, retention in the primary grades had a positive effect on academic self-efficacy in fourth grade (Reynolds, 1992), but this effect disappeared by the age of 14 years (McCoy & Reynolds, 1999), suggesting that the short-term and longer term effects of early grade retention on self-efficacy may differ.

The Present Study

Drawing on social comparison theory in general and the big fish, little pond effect more specifically, we expected that retained students would show improvements in psychosocial adjustment across the board during their repeat year, relative to their propensity score–matched peers. However, contrary to the view that the enhanced social and academic performance during the repeat year will be a turning point for retained students (Tomchin & Impara 1992), we hypothesized that retained children's short-term gains will diminish over the span of 2 to 4 years as they struggle in meeting new academic challenges and classroom social comparison cues no longer favor them. In other words, we expect that the "gift of time" has an expiration date for at least some aspects of psychosocial adjustment and peer relations. These expectations are informed by two studies with this same longitudinal data set. First, in a study investigating only academic outcomes and using piecewise growth modeling, Wu et al. (2008a) found that retention led to improved scores relative to same-grade peers on measures of math and reading achievement, as measured by Woodcock–Johnson grade standard scores (Woodcock & Johnson, 1989; Woodcock, McGrew, & Mather, 2001) in the short term (Years 1–2). However, these short-term gains in grade-level math and reading diminished over the next 2 years as retained students encountered novel curricula and routines. Second, Gleason et al. (2007) found that retained students, relative to same-age promoted peers, were better accepted by their classmates during their repeat year (Year 2), and that this effect was completely mediated by teacher and peer perceptions of students' academic competence during Year 2. The current study extends our prior work by investigating multiple dimensions of psychosocial adjustment over 4 years using stronger controls for selection effects than were used in the Gleason et al. study.

We used two primary approaches to address the issue of the baseline comparability of retained and promoted children that has plagued previous studies (Lorence, 2006). First, prior to any of the children being retained, we selected and comprehensively assessed a sample of children who had scores below the 50th percentile on school district tests of reading. This approach allowed us to focus on the group of children who were at some risk of being retained given the known relationship between retention and poor reading skills. Second, we used propensity score matching, which corrects for selection bias associated with all measured covariates (Rosenbaum, 2002; West & Thoemmes, in press). This approach allowed us to closely match retained and nonretained children on the basis of an extensive, carefully chosen set of variables collected at baseline. Taken together, these approaches allowed us to estimate a meaningful treatment effect—the causal effect of retention for the treated, that is, those children who were actually retained in first grade (see Schafer & Kang, 2008). In research contexts in which a large proportion of the population (here, the higher achieving children) have no probability of receiving the treatment, the estimate of the causal effect should be based only on those children who could potentially receive the treatment (Rubin, 2005).

For our primary analyses, we used growth curve modeling that permits estimation of each child's trajectory of growth on teacher and peer reports of students' internalizing and externalizing behaviors, peer reports of acceptance, and teacher reports of student effortful engagement in the classroom. We measured each of these variables using the identical measurement instrument at each measurement wave, meeting the requirements for growth curve modeling. With four waves of data, we could examine the effects of retention on both the short-term and longer term growth on each outcome, using piecewise linear trajectory models (Singer & Willett, 2003). This approach allowed us to compare the differences in the growth of the retained and promoted children in the short term (Year 1 through Year 2), when retained children are repeating the first-grade curriculum, as well as in subsequent years (Year 2 through Year 4), when retained and promoted children are exposed to novel curricula at new grade levels. In accordance with our hypotheses, we expected that short- and longer term effects would differ on several of the variables. We collected two additional relevant psychosocial variables (i.e., children's perceived self-concept and school belonging), but consistent with the current recommendations of the test developers, assessed these using different, developmentally appropriate measures over the 4 years of the study. Because these measures do not share a common metric across the 4 years of the study, growth curve modeling was not possible (Khoo, West, Wu, & Kwok, 2006). For these measures, we were only able to compare the levels of the retained and promoted children at each time point.

Method

Participants

Participants were drawn from a larger sample of children participating in a longitudinal study examining the impact of grade retention. Participants were recruited from three school districts in Texas (one urban district and two small city districts) across two sequential cohorts in first grade during the fall of 2001 and 2002. The composition of the urban school district was 41% White non-Hispanic, 37% economically disadvantaged, and 11% limited English proficient. Enrollment in one of the small city school districts was 40% White non-Hispanic, 61% economically disadvantaged, and 11% limited English proficient. The enrollment of the second small city school district was 69% White non-Hispanic, 24% economically disadvantaged, and 5.2% limited English proficient. Children were eligible to participate in the longitudinal study if they scored below the median on a state-approved district-administered measure of literacy, spoke either English or Spanish, were not receiving special education services, and had not been previously retained in first grade. Teachers were asked to distribute consent forms to parents of all 1,374 eligible children via children's weekly folders; thus the exact number of parents who actually received the consent forms could not be determined. Teachers and parents were told that the purpose of the study was to learn more about factors that influence children's adjustment to and success in school. Incentives for returning the consent forms, regardless of whether consent was granted, in the form of small gifts to children (e.g., erasers, fancy pencils) and the opportunity to win a larger prize in a random drawing were instrumental in obtaining a return of 1,200 consent forms placed in children's folders. A total of 784 parents (65%) provided consent, and 416 declined. The research was approved by the Institutional Review Boards of Texas A&M University, Arizona State University, and each school district's research advisory team.

Analyses of a broad array of archival variables—including performance on the district-administered test of literacy (standardized within district as a result of differences in test used), age, gender, ethnicity, eligibility for free or reduced-price lunch, bilingual class placement, cohort, and school context variables (i.e., percentage ethnic/racial minority; percentage economically disadvantaged—did not indicate any differences between children with and without consent. The resulting sample of 784 participants (52.6% male, 47.4% female) closely resembles the population from which the participants 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% of the children 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).

Participants for the growth curve analysis consisted of the 375 children (33% retained, 53% male) who were successfully matched with respect to their propensity to be retained in first grade (see description of propensity score matching procedure below). The racial/ethnic composition of the sample was 35% Caucasian, 34% Hispanic, 27% African American, and 4% other. For each outcome, we only included those participants who had observations on that outcome for at least one of the four assessment periods in the growth curve modeling. The number of retained and promoted children who were included in the analysis for each of the outcomes is shown in Table 1. The mean correlations among the 10 outcome measures across measurement waves are presented in Table 2.

Overview of Design and Measures

Demographic information, including child age, gender, race/ethnicity, eligibility for free or reduced lunch, and status as limited English proficient (LEP), was obtained from school district records. Teacher and peer data were collected annually beginning when all participants were in first grade.[1] Teachers and parents received $25 for completing and returning the questionnaires. Peers' perceptions of the level of externalizing behaviors were obtained via individual interviews. Children's perceived self-efficacy and sense of school belonging were also obtained in individual interviews. To minimize language factors in children's responses, LEP children, children in bilingual classrooms, and children with Hispanic surnames were first administered a language proficiency test by a Spanish–English bilingual examiner to establish the child's dominant language, and all tests were administered in that language.

A total of 72 baseline variables were collected at measurement Wave 1 and used in the calculation of propensity scores. The baseline variables included demographic measures, cognitive and behavioral performance, social and emotional functioning, and classroom and school variables. The 72 variables were intended to be as comprehensive as possible, including variables that have been shown in prior research to be related to early retention versus promotion, outcomes, or ideally both.[2]

A total of eight teacher and peer report measures of psychosocial outcomes served as the primary outcome measures for this study, as shown in Table 1. These measures were administered on an annual basis to teachers and peers (described below). Two additional child report outcomes were assessed with age-appropriate measures at each measurement wave. The descriptive statistics for the eight teacher- and peer-reported and two child-

[1] Very few students had the same teacher across years. Among the 375 participants in the study, only 13 students (3.5%) had the same teacher at Time 1 and 2 measurement waves, 5 students (1.3%) had the same teacher at Times 2 and 3, and 3 students (0.8%) had the same teacher at Times 3 and 4. None of the students had the same teacher over all 3 years.

[2] A complete list of the 72 variables collected at Wave 1 is available from Jan N. Hughes.

Table 1
Descriptive Statistics for 10 Psychosocial Outcomes

Variable	No. of retained children[a]	No. of promoted children[b]	Means for retained children				Means for promoted children			
			Grade 1	Grade 2	Grade 3	Grade 4	Grade 1	Grade 2	Grade 3	Grade 4
Behavioral adjustment										
Externalizing behaviors										
Teacher-reported hyperactivity	122	241	1.01 (0.65)	0.88 (0.67)	0.84 (0.66)	0.82 (0.69)	0.93 (0.63)	0.87 (0.61)	1.00 (0.67)	0.83 (0.62)
Teacher-reported conduct problems	122	240	0.36 (0.46)	0.46 (0.55)	0.36 (0.43)	0.34 (0.45)	0.40 (0.50)	0.40 (0.50)	0.39 (0.50)	0.37 (0.49)
Peer-reported hyperactivity	111	211	0.22 (1.06)	0.14 (1.04)	0.26 (1.14)	0.11 (0.93)	0.10 (0.99)	0.07 (1.01)	0.05 (0.91)	0.03 (0.93)
Peer-reported conduct problems	111	213	0.09 (1.01)	0.22 (1.11)	0.26 (1.16)	0.05 (1.00)	0.05 (1.01)	0.05 (1.00)	0.09 (1.07)	0.04 (0.92)
Internalizing behaviors										
Teacher-reported emotional problems	122	240	0.39 (0.43)	0.37 (0.41)	0.29 (0.39)	0.33 (0.36)	0.41 (0.43)	0.37 (0.41)	0.38 (0.40)	0.43 (0.45)
Peer-reported sad/withdrawn	111	213	0.19 (1.08)	−0.22 (0.66)	−0.06 (0.90)	−0.17 (0.75)	0.06 (1.02)	0.001 (0.93)	0.09 (1.10)	−0.01 (0.88)
Engagement										
Teacher-reported behavioral engagement	123	243	2.92 (1.04)	3.29 (1.03)	3.39 (1.08)	3.39 (0.76)	03.15 (1.06)	3.22 (1.04)	3.06 (0.95)	3.34 (0.78)
Child-reported school belonging[c]			4.13 (0.80)	4.23 (0.68)	4.21 (0.79)	3.92 (0.69)	4.19 (0.83)	4.03 (0.78)	4.09 (0.69)	3.81 (0.67)
Child-reported academic self-efficacy[c]			3.44 (0.56)	3.55 (0.39)	22.98 (5.53)	22.71 (5.69)	3.42 (0.60)	3.44 (0.46)	21.61 (5.02)	21.23 (4.63)
Social acceptance										
Peer-reported liking	111	221	3.40 (0.72)	3.60 (0.68)	3.37 (0.60)	3.13 (0.69)	3.38 (0.71)	3.15 (0.67)	3.09 (0.63)	3.12 (0.59)

Note. Standard deviations are presented within parentheses.
[a] Number of retained children used in the analysis. [b] Number of promoted children used in the analysis. Sample sizes vary slightly as a result of missing data. [c] Descriptive statistics for school belonging and academic self-efficacy before imputation. The number of retained and promoted children varied across time. Included in the analysis were 125, 103, 98, and 97 retained children and 252, 218, 224, and 212 promoted children at Grades 1, 2, 3, and 4, respectively.

reported outcomes for retained and promoted children at each of the four time points are also shown in Table 1.

Measures

Externalizing behaviors.

Teacher-reported hyperactivity and conduct problems. Teachers completed the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997, 2001), a brief (25-item) screening measure for psychopathology. Each item is rated on a three-point scale (0 = *not true*; 1 = *somewhat true*; 2 = *certainly true*). The SDQ yields five scales composed of 5 items each. Coefficient alpha for the Conduct Problem scale ranged from .82 to .83 across the four waves. Coefficient alpha for the Hyperactivity scale ranged from .86 to .88. One-year test–retest correlations (stability coefficients) for adjacent measurement waves ranged from .57 to .67 for conduct problems and from .57 to .61 for hyperactivity. In a sample of children participating in this longitudinal study, teacher reports of conduct problems and hyperactivity were moderately to strongly correlated with both parent and peer reports (Hill & Hughes, 2007). Factor analyses support the construct validity of the SDQ (Dickey & Blumberg, 2004; Hill & Hughes, 2007).

Peer-reported hyperactivity and conduct problems. We followed procedures widely recommended in the peer assessment literature (Cillessen & Bukowski, 2000) to assess peers' perceptions of classmates' hyperactivity and aggression. In individual interviews, students were presented with a roster containing the names of all classmates. The interviewer read all classmates' names and asked the child if he or she knew each child. The interviewer then asked children to nominate as few or as many classmates as they wished who fit each descriptor. The aggression item read: "Some kids start fights, say mean things, or hit others." The hyperactivity item read: "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 score and a hyperactivity score based on the number of nominations that the child received. Stability coefficients across the 4 years ranged from .45 to .52 for aggression and from .47 to .48 for hyperactivity. Sociometric scores were standardized within classrooms. Written parental 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. The mean rate of classmate participation in the sociometric administrations was .65 (range = .40 to .95). Elementary school children's peer nomination scores derived from procedures similar to those used in this study have been found to be reliable and stable over periods of 6 weeks to 4 years and to be associated with concurrent and future behavior and adjustment (Cillessen & Bukowski, 2000; Realmuto, August, Sieler, & Pessoa-Brandao, 1997).

Internalizing problems.

Teacher-rated emotional problems. The five-item Emotional Problems scale of the Strengths and Difficulties Questionnaire (Goodman, 1997), described above, assessed children's sad and withdrawn behaviors. Across the 4 years, coefficient alpha ranged from .69 to .77. Scores correlated significantly with parent and peer reports of emotional problems, and the scale was supported by confirmatory factor analyses (Hill & Hughes, 2007).

Peer-rated emotional problems. Using peer nomination procedures described above, classmates were asked to nominate chil-

Table 2
Mean Correlations Among the Outcome Measures Across Measurement Waves

Variable	1	2	3	4	5	6	7	8	9	10
Behavioral adjustment										
Externalizing behaviors										
1. Teacher-reported hyperactivity	—									
2. Teacher-reported conduct problems	.59**	—								
3. Peer-reported hyperactivity	.49**	.47**	—							
4. Peer-reported conduct problems	.39**	.52**	.64**	—						
Internalizing behaviors										
5. Teacher-reported emotional problems	.31**	.29**	.08	.09	—					
6. Peer-reported sad/withdrawn	.06	.07	.10*	.08	.19**	—				
Engagement										
7. Teacher-reported behavioral engagement	−.77**	−.53**	.32**	−.29**	−.33**	−.02	—			
8. Child-reported school belonging	−.18**	−.09	−.09	−.09	−.06	−.01	.15**	—		
9. Child-reported academic self-efficacy	−.08	.02	.02	.06	−.09	−.03	.02	.25**	—	
Social Acceptance										
10. Peer-reported Liking	−.34**	−.28**	−.32**	−.29**	−.10*	.11*	.31**	.10*	.05	—

Note. Correlations are based on the pairwise deleted correlation matrix for the full at-risk sample. *N*s ranged from 368 to 733 because of missing data. On the basis of the minimum *N* of 368, $r = .10$ is significant at $p = .05$, and $r = .12$ is significant at $p = .01$.
$^{*}p < .05$. $^{**}p < .01$.

dren who met the following description: “Some kids cry a lot and look sad.” Nominations were summed and standardized within classrooms. Stability coefficients across the 4 years ranged from .33 to .42. The lower stability coefficients for sad/withdrawn behaviors likely reflect the greater difficulty that peers have in discerning more covert constructs compared with more observable constructs (Funder & West, 1993).

Behavioral engagement. Teachers completed a 10-item Likert-type scale. Example items are “Is a reliable worker”; “Perseveres until the task is finished”; “Tends to be lazy” (reverse scored), “Is easily distracted”; and “Sets and works toward goals.” Coefficient alpha for the scale ranged from .91 to .95 across the four measurement waves. The stability coefficients for teacher-rated behavioral engagement ranged from .55 to .61. In a sample of children participating in this longitudinal study, teacher-rated engagement on this measure predicted cross-year academic achievement, above the effects of prior achievement, IQ, antisocial engagement, and family background variables (Hughes et al., 2008).

Affective engagement: Sense of school belonging. On the basis of developmental considerations, the measure of school belonging used in Years 1–3 (the initial primary grades) differed from that introduced at Year 4. In Years 1–3, an experimenter-developed measure was used to evaluate children’s perception of liking for and sense of belonging at school. In individual interviews, children were asked to indicate how they felt in response to five statements about school by pointing to one of five faces whose expressions ranged from *very sad* (1) to *very happy* (5). Example items included “How much does your teacher enjoy spending time with you?”; “How much do you like to go to school?”; and “How do you feel when you are at school?” Coefficient alpha values for the scale were .60, .61, and .67 at Grades 1, 2, and 3 for the sample. In Year 4, students were administered the Psychological Sense of School Membership (Goodenow, 1993) instrument, which asks students to indicate their agreement, on a 5-point Likert-type scale, with 18 items that assess students’ perceived acceptance, feelings of inclusion, and encouragement for participation (Goodenow, 1993). The scale has good evidence of validity (Goodenow, 1993; Hagborg, 1998). Coefficient alpha for the scale was .83.

Cognitive engagement: Perceived academic self-efficacy. In Years 1 and 2, we used the Pictorial Scale of Perceived Competence and Social Acceptance for Young Children (PSPCSA; Harter & Pike, 1981) to assess children’s academic self-efficacy beliefs. The Cognitive Competence subscale consists of six items. For each item, children are presented with pictures of two children who are described in contrasting ways (e.g., “This girl is good at spelling”; “This girl is not good at spelling”). Children are then asked which child is more like them. After making their choice, children are asked if that child is a little or a lot like them. The procedure yields a 4-point scale for each item. Example items are “good at numbers,” “knows a lot in school,” and “can read alone.” Coefficient alpha for the scale for our sample was .77 and .71 at Years 1 and 2, respectively. In Years 3 and 4, the children’s perceived reading and math competencies were assessed with the Competence Beliefs and Subjective Task Values Questionnaire (Wigfield et al., 1997). Children were asked how good they are in that domain, how good they are relative to the other things they do, how good they are relative to other children, how well they expected to do in the future in that domain, and how good they thought they would be at learning something new in that domain. Children were asked to indicate their response by pointing to a 0- to 30-point thermometer-type scale, with 0 being the lowest and 30 being the highest. This questionnaire has strong evidence of reliability and validity with students as young as third graders (Wigfield et al., 1997). Coefficient alpha for our sample was .82 and .84 at Years 3 and 4, respectively.

Social acceptance. Children also were asked to indicate their liking for each child in the classroom on a 5-point scale. The interviewer named each child in the classroom and asked the child to point to one of five faces ranging from sad (1 = *don’t like at all*) to happy (5 = *like very much*). A child’s mean liking score was the average rating received by classmates. Scores were standardized within classrooms. Ratings of peer liking are relatively stable over the elementary school years and are associated with a number of indices of behavioral adjustment (Bierman, 2004; Hughes, 1990). The stability coefficient for the mean liking score ranged from .45 to .52 across the four measurement waves.

Propensity Score Estimation

Propensity scores, the predicted probability of being retained in first grade, were estimated for the full sample of 768 children for whom retention information was available. We used 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 logistic regression equation led to relatively good prediction of the decision to retain or promote each child, with a Nagelkerke pseudo R^2 index of .552 (see Cohen, Cohen, West, & Aiken, 2003, p. 503).

The propensity score can range from 0 to 1. The larger the propensity score, the higher the probability that the child would be retained in the first grade. For the 768 children who were below the median on reading at school entrance, the propensity score ranged from .0003 to .989, with M = .215 and SD = .215. In this full sample of at-risk children, the children who were subsequently promoted had substantially lower propensity scores (N = 603; M = .126, SD = .163) than did those who were subsequently retained (N = 165; M = .540, SD = .292), $t(766) = -23.82$; Cohen's $d = -2.09$, $p < .001$. Figure 1A shows separate kernel density estimates of the distribution of propensity scores for promoted and retained children for the 768 cases. Kernel density estimates smooth the data, providing an estimate of the distributions for the retained and promoted children in the population (Cohen et al., 2003, pp. 105–108). The figure shows that the distribution of propensity scores for the promoted children was highly right-skewed, whereas the distribution for the retained children was relatively uniform across the full range of propensity scores.

Despite identifying an at-risk sample of children who were below the median on literacy at entrance to first grade, we found substantial differences between the retained and promoted groups. These results indicated that an adjustment procedure would be needed to equate the retained and promoted groups. Following the recommendations of Rosenbaum (2002), we chose a procedure that produces optimal matches on propensity scores.

Matching Procedure

We matched retained and promoted children on the basis of their propensity scores. Rather than conducting a 1:1 matching between retained and promoted children, we chose to perform a 1:many matching, with a variable number of matches between retained and promoted children. Ming and Rosenbaum (2000) showed that this procedure can remove substantially more bias than fixed ratio matching and maximize sample size at the same time. Over the range of propensity scores from .00 to .50, there were more promoted children than retained children in the sample. For this range, we matched 1 retained child with up to 5 (i.e., 1, 2, 3, 4, or 5) promoted children. Over the range of propensity scores from .50 to 1.00, there were more retained children than promoted children in the sample. For this range, we matched each promoted child with up to 5 retained children. Otherwise stated, a target child was selected from the smaller group (retained group for a propensity score of <.50; promoted group for a propensity score of ≥.50) and was then matched with up to 5 children from the other, larger group. To assure high-quality matching, we imposed a caliper distance of .025, the maximum distance in propensity scores that was allowed for a match to take place. That is, any pair of retained and promoted children who differed in their propensity scores by more than .025 could not be matched with each other.

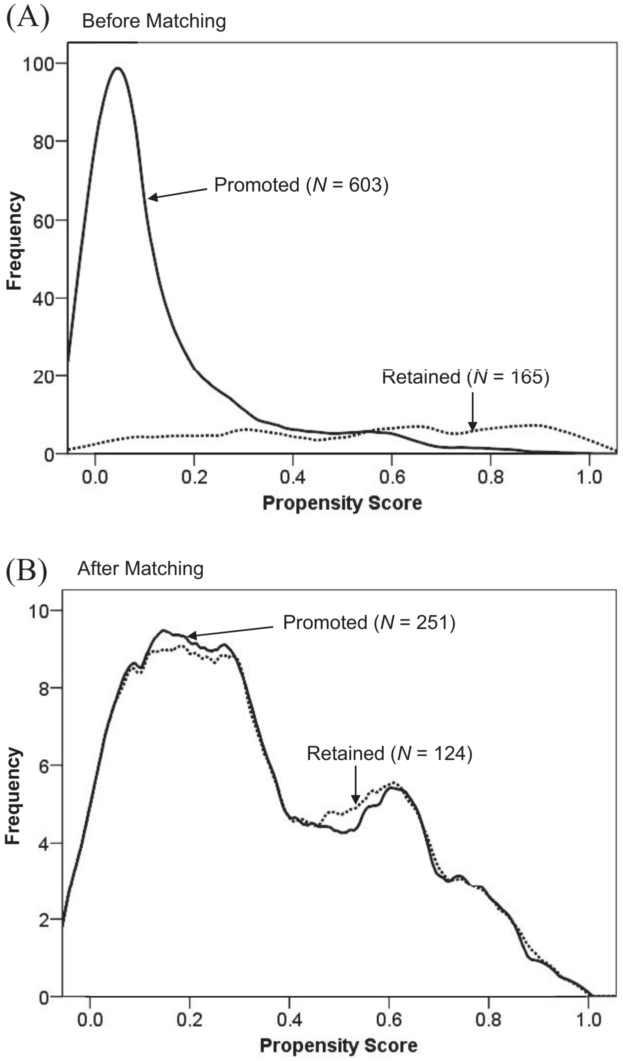


Figure 1. Kernel density plots of (A) the frequency distributions of the propensity score for promoted and retained children before optimal matching and (B) the frequency distributions of the mean propensity score for matched sets of promoted and retained children after optimal matching. The scale of the Y-axes differs between the two panels.

SAS 8.0 PROC ASSIGN was used to implement the matching (Ming & Rosenbaum, 2001).[3] PROC ASSIGN matches retained children with promoted children so that the sum of the distances between the propensity scores within each of the matched sets was minimized for the whole sample. A total of 80 matched sets were constructed with a total of 251 promoted and 124 retained children. For the 80 matched sets, the propensity score ranged from .003 to .934, with $M = .31$ and $SD = .25$. Our empirical range covered virtually the entire theoretical range for the propensity scores (.00 to 1.00). As is shown by the kernel density estimates in Figure 1B, the two groups were closely equated on their propensity scores following the optimal matching process, with the two distributions almost overlapping each other. This result indicates that good balance between the two groups was attained on the propensity scores. Following the recommendation of Rubin (2006), the matching procedures were conducted without any knowledge of the participant's scores on any variable collected after baseline.

To further test whether the matching provides good balance between the retained and promoted groups, we compared the retained and promoted groups on all baseline measures used to calculate the propensity score and the baseline measures for all of the outcomes examined in the study, which includes 75 baseline measures in total. Given space limitations, we report balance for 20 important baseline measures. We divided the propensity scores into five strata (quintile groups): 0–19th percentile; 20–39th percentile, 40th–59th percentile, 60th–79th percentile, and 80th–100th percentile. Next, we tested whether the effect of grade retention on the 20 baseline measures differed from 0 for each of the five groups. For the continuous variables, we conducted a 2 (retained vs. promoted) × 5 (quintile) analysis of variance (ANOVA) using the baseline measures (Table 3). For dichotomous variables, we conducted a parallel 2 (retained vs. promoted) × 5 (quintile) analysis using logistic regression (Table 4). If a baseline measure is well balanced between the retained and promoted groups, then neither the main effect nor the interaction should be different from 0. Tables 3 and 4 show our matching procedure provided good balance on 18 of the 20 baseline measures.[4] The only exceptions among the 20 important baseline variables were the Woodcock–Johnson broad math achievement score and White vs. non-White ethnicity.

Data Analysis

Growth curve modeling for teacher and peer report measures. To investigate the relatively short-term and longer term effects of retention on the change of the teacher and peer report measures, we split the time span into two pieces at the point where measurement wave = 2 (the 2nd year of the study). Thus, short-term change compares the change from Year 1 to Year 2, when the retained children are repeating the first-grade curriculum but the promoted students are encountering new curriculum. Longer term change examines growth from Years 2 through 4, when both the retained and promoted children are encountering new curricula. We fit a linear growth curve on each piece for the eight psychosocial outcomes separately using SAS 8.0 PROC Mixed. Such growth curve models are termed two-piece linear growth curve models (Singer & Willett, 2003). We used full information maximum likelihood estimation (FIML) to estimate the growth curve models, given the existence of missing data. FIML utilizes all of the observations available for each case to compute the likelihood function (Enders & Bandalos, 2001). FIML provides unbiased estimates with minimal standard errors when data are missing at random (Schafer & Graham, 2002). Otherwise stated, FIML provides estimates that are appropriately corrected for all measured variables included in the analysis.

The specification of the two-piece linear growth curve model for each outcome is shown in Equations 1 to 3 below (see p. 145). The model includes three levels. Level 1 (within individual; Equation 1) estimates the two-piece linear growth trajectory for each individual over the 4 years of the study. At Level 1, two time variables ($T1_{tip}$ and $T2_{tip}$), corresponding to the two pieces, were coded to predict the outcome (see Table 5). This coding scheme permitted us to test the slope separately for each piece and, further, to test whether the slopes in each piece differ between retained and promoted groups. We centered time at Wave 1 (the point at which both $T1_{tip}$ and $T2_{tip}$ were coded 0); thus the intercept represents the student's initial status on the 10 outcomes in first grade prior to retention or promotion.

Level 2 (between individuals; Equation 2) captured 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 individual intercepts, individual slopes for the short term (slope 1) and individual slopes for the longer term (slope 2). Because the data for the children within the same many-to-one matched set may have higher correlation than those in different matched sets (dependency caused by matching), we added Level 3 (between matched set; Equation 3) to take into account the within matched set correlation for the individual intercepts. 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 effects of retention on slope 1 (short-term effect, γ_{110}) and slope 2 (longer term effect, γ_{210}) are of primary interest. The equations for the three-level model for each of the outcomes are presented below.

[3] We tried both 1:up to 3 and 1:up to 5 matching with caliper distance = .05 and .025. We found that 1:up to 5 matching with caliper distance = .025 led to the smallest total within-match distance without much sacrifice of the number of matched cases compared with the other conditions. Thus 1:up to 5 matching with caliper distance = .025 was adopted in the study.

[4] Our many-to-one matching procedure led to overall good balance on 75 baseline variables. A series of 2 (retention) × 5 (quantile strata) ANOVAs for continuous variables and logistic regressions for dichotomous variables identified six significant effects at $p < .05$, whereas 7.5 would be expected by chance. The maximum effect size on the baseline measures ($\eta^2 = .047$) was less than moderate in magnitude according to Cohen's (1988) guidelines. Significant baseline main effects or interactions involving retention were found on math raw achievement, ethnicity (White vs. non-White), parent-rated internalizing problems, percentage of White students in class, and family adversity among the 75 baseline measures. We conducted sensitivity analyses using (a) the five significant measures from the full set of 75 baseline measures and (b) the two significant measures from the set of 20 important baseline measures as covariates in the Level 2 model to adjust for baseline differences. The effects of retention after partialing out these sets of covariates did not differ materially from those without the covariates added.

Table 3

Checks on the Success of Variable Many-to-One Matching. Continuous Variables: F Tests From Analyses of Variance

Variable	Main effect of retention: *F*(1, 251)	Retention × Quintile Strata: *F*(4, 251)
Behavioral adjustment		
Externalizing behaviors		
Teacher-reported hyperactivity	0.05	0.07
Teacher-reported conduct problems	0.35	0.53
Peer-reported hyperactivity	0.14	1.38
Peer-reported conduct problems	0.05	2.08
Internalizing behaviors		
Teacher-reported emotional problems	0.55	1.86
Peer-reported sad/withdrawn	2.56	0.55
Engagement		
Teacher-reported behavioral engagement	0.01	0.82
Child-reported school belonging	1.47	0.18
Child-reported academic self-efficacy	0.01	0.05
Social acceptance		
Peer-reported liking	1.30	0.29
Other measures		
Child age	0.08	0.64
Child IQ	0.01	1.07
Parent highest level of education	0.85	0.93
Parent highest level of employment	0.87	0.39
Child Woodcock–Johnson Math W score	0.20	2.68*
Child Woodcock–Johnson Reading W score	1.05	1.34

Note. Tables 3 and 4 report the 40 tests for the 20 important variables, of which 3 were statistically significant. We would expect 2 of the total number of tests to be significant by chance.

* $p < .05$.

Level 1:

$$Y_{tip} = \pi_{0ip} + \pi_{1ip}T1_{tip} + \pi_{2ip}T2_{tip} + e_{tip};\ e_{tip} \sim N(0, \sigma^2). \quad (1)$$

Level 2:

$$\pi_{0ip} = \beta_{00p} + \beta_{01p}RETENTION_{ip} + r_{0ip};$$

$$\pi_{1ip} = \beta_{10p} + \beta_{11p}RETENTION_{ip};$$

$$\pi_{2ip} = \beta_{20p} + \beta_{21p}RETENTION_{ip};\ r_{0ip} \sim N(0, \tau_{\pi 00}). \quad (2)$$

Level 3:

$$\beta_{00p} = \gamma_{000} + u_{00p};\ \beta_{10p} = \gamma_{100};\ \beta_{20p} = \gamma_{200};$$

$$\beta_{01p} = \gamma_{010};\ \beta_{11p} = \gamma_{110};\ \beta_{21p} = \gamma_{210};$$

$$u_{00p} \sim N(0, \tau_{\beta 00}). \quad (3)$$

Table 4

Checks on the Success of Variable Many-to-One Matching—Binary Variables: Wald Tests From Logistic Regression

	Logistic regression	
Baseline measure	Main effect of retention (Wald χ^2, *df* = 1)	Retention × Quintile strata (Wald χ^2, *df* = 1)
Child ethnicity: White vs. non-White[a]	5.40*	7.25*
Child gender	1.20	0.52
Child bilingual status[b]	0.13	0.001
Child economic disadvantage status[b]	0.03	0.16

Note. Tables 3 and 4 report the 40 tests for the 20 important variables, of which 3 were statistically significant. We would expect 2 of the total number of tests to be significant by chance.

[a] Non-White ethnicities included African American, Hispanic, Asian/Pacific Islander, Native American/Alaskan, and other. [b] 1 = yes; 0 = no.

* $p < .05$.

Table 5
Coding Scheme

Time variable	Wave 1	Wave 2	Wave 3	Wave 4
$T1$	0	1	1	1
$T2$	0	0	1	2

Here subscript t indicates time point (Wave 1, 2, 3, or 4), subscript i indicates individual, and subscript p indicates matched set. Y_{tip} is the outcome. γ_{000} and γ_{100} represent the grand mean of the intercept and slope 1, respectively, for the promoted group. γ_{200} represents the grand mean of slope 2 for the promoted group. γ_{010} and γ_{110} represent the estimated effects of grade retention in first grade on the intercept and slope 1, respectively. γ_{210} represents the effect of grade retention on slope 2. e_{tip} represents Level 1 residual for the ith individual within the pth set at wave t, which was assumed to follow a normal distribution with mean $(\mu) = 0$ and homogeneous variance (σ^2) across 4 years. σ^2 is the within-individual variance in Woodcock–Johnson scores that cannot be accounted for by time. r_{0ip} represents the Level 2 residual in the intercept for the ith individual within the pth matched set, which was assumed to follow a normal distribution, with $\mu = 0$ and variance $= \tau_{\pi 00}$. $\tau_{\pi 00}$ is the between-individual variance in the intercept that cannot be accounted for by grade retention. u_{00p} represents the deviation of the mean intercept of the pth set from the grand mean intercept for promoted children, which was also assumed to follow a normal distribution with $\mu = 0$ and variance $= \tau_{\beta 00}$. $\tau_{\beta 00}$ is the between matched set variance in intercept. After including grade retention as a predictor at the individual level (Level 2), we found that residual variances for all of the outcomes were not different from 0. There was also no substantial variance in slopes at Level 3. Thus, in the growth curve models shown in Equations 1, 2 and 3, we constrained the residual variances in slopes at Level 2 and the variances in slopes at Level 3 to 0.

Regression analyses for child report measures. The measures of academic self-efficacy and school belonging shifted over the four measurement waves; we could not conduct growth curve models on these data. Instead, we conducted a series of regression analyses to examine the effect of grade retention on these measures at each measurement wave following retention (Grades 2 to 4). In the regression model, we also included the baseline measure of the outcome variable and the propensity score (probability of being retained at the end of Grade 1) as predictors to control for any difference between the retained and promoted groups prior to retention. Given that there were missing observations for the two outcome variables at Grades 2 to 4, we followed the recommendation of Schafer and Graham (2002) and used multiple imputation (MI) to impute values for the missing data. We generated 50 complete data sets using Proc MI in SAS 9.0. In the imputation model, we included 80 auxiliary variables that had $>.30$ correlations with the missingness of either of the outcome variables at Grades 2, 3, or 4. The use of auxiliary variables follows the recommendations of Collins, Schafer, and Kam (2001) that the performance of MI can be improved substantially through the use of auxiliary variables that are correlated with missingness. We then ran the regression analyses on each of the 50 data sets and combined the parameter estimates from the 50 runs using PROC MIANALYZE (see Schafer & Graham, 2002).

Results

Growth Curve Modeling of the Teacher and Peer Report Measures

Parameter estimates. Table 6 presents the estimated intercept, linear growth rate in the short term (slope 1), and growth rate in the longer term (slope 2) for both promoted and retained children as well as the effect of retention in the first grade on the intercept, slope 1, and slope 2, which indicate the differences in intercept, slope 1, and slope 2 between retained and promoted children, for each of the teacher and peer report measures. We use $S1_P$ and $S2_P$ to represent slope 1 and slope 2 for promoted children and $S1_R$ and $S2_R$ to represent slope 1 and slope 2 for retained children.

Retention led to short- or longer term effects on four of the eight outcomes. Retention had a negative effect on short-term slope for teacher-rated hyperactivity ($S1_R - S1_P = -.18$, with $SD = .08$; $d = -.11$, $p < .05$)[5] and peer-rated sad/withdrawn behaviors ($S1_R - S1_P = -.34$, with $SD = .14$; $d = -.34$, $p < .05$). Retained children experienced a decrease (i.e., improvement) in both teacher-rated hyperactivity ($S1_R = -.20$, with $SD = .06$; $d = -.19$, $p < .01$) and peer-rated sad/withdrawn behaviors ($S1_R = -.37$, with $SD = .12$; $d = -.33$, $p < .01$) in the short term (see Figures 2A and 2B). In contrast, promoted children experienced little change in either teacher-rated hyperactivity or peer-rated sad/withdrawn behaviors in the short term. Retention had a positive effect on short-term slope for behavioral engagement ($S1_R - S1_P = .43$, with $SD = .12$; $d = .27$, $p < .01$). Retained children showed an increase (i.e., improvement) in behavioral engagement ($S1_R = .46$, with $SD = .10$; $d = .35$, $p < .01$) in the short term, whereas promoted children showed little change in behavioral engagement in the short term (see Figure 2C). In the longer term, the retained and promoted children did not differ in their slopes for teacher-rated hyperactivity, peer-rated sad/withdrawn behaviors, or behavioral engagement. This result indicates that the short-term improvements were maintained for these three outcomes.

In addition, retention had a positive effect on short-term slope ($S1_R - S1_P = .46$, with $SD = .09$; $d = .60$, $p < .01$) and a negative effect on longer term slope ($S2_R - S2_P = -.20$, with $SD = .05$; $d = -.59$, $p < .01$) for peer-rated liking. Retained children experienced an increase in mean peer-rated liking in the short term ($S1_R = .21$, with $SD = .07$; $d = .27$, $p < .01$) but a negative slope for mean peer-rated liking ($S2_R = -.23$, with $SD = .04$; $d = -.67$, $p < .01$) in the longer term (see Figure 2D). In contrast, for promoted children, there was a decrease in mean peer-rated liking in the short term ($S1_P = -.25$, with $SD = .05$; $d = -.32$, $p < .01$) and no change in mean peer-rated liking in the longer term. The immediate gains in peer-rated liking for retained relative to promoted children appear to dissipate within a few years after retention.

[5] The variable d represents the effect size measure. According to Cohen (1988), d values of .20, .50, and .80 represent small, medium, and large effect sizes, respectively.

Table 6

Effects of Retention at Grade 1 on Growth Parameters in the Piecewise Model

Variable	Means for promoted children			Means for retained children			Effect of retention on			Model fit
	Intercept	Slope 1	Slope 2	Intercept	Slope 1	Slope 2	Intercept	Slope 1	Slope 2	Conditional concordance correlation
Behavioral adjustment										
Externalizing behaviors										
Teacher-reported hyperactivity	0.94* (0.05)	−.02 (.04)	−.01 (.08)	1.04* (0.06)	−.20* (.06)	−.01 (.03)	0.10 (0.08)	−.18* (0.08)	−.001 (.04)	0.83
Teacher-reported conduct problems	0.41* (0.04)	−.01 (.03)	−.004 (.02)	0.36* (0.05)	.07 (.05)	−.04 (.06)	−0.03 (0.06)	.08 (.06)	−.04 (.03)	0.86
Peer-reported hyperactivity	0.09 (0.07)	−.01 (.08)	−.02 (.04)	0.18 (0.10)	−.02 (.11)	.004 (.06)	0.09 (0.13)	−.01 (.13)	.02 (.07)	0.78
Peer-reported conduct problems	0.04 (0.08)	.03 (.08)	−.02 (.04)	0.06 (0.11)	.19 (.11)	−.06 (.06)	0.02 (0.13)	.16 (.13)	−.05 (.07)	0.82
Internalizing behaviors										
Teacher-reported emotional problems	0.41* (0.03)	−.05 (.04)	.04* (.02)	0.39* (0.04)	−.05 (.05)	−.02 (.03)	−0.02 (0.05)	.01 (.07)	−.06 (.04)	0.31
Peer-reported sad/withdrawn	0.04 (0.07)	−.03 (.08)	.02 (.05)	0.20* (0.10)	−.37* (.12)	.03 (.07)	0.16 (0.12)	−.34* (.14)	.02 (.08)	0.67
Engagement										
Teacher-reported behavioral engagement	3.11* (0.07)	.03 (.07)	.05 (.04)	2.90* (0.10)	.46* (.10)	.02 (.05)	−0.21 (0.12)	.43* (.12)	−.02 (.07)	0.78
Social acceptance										
Peer-reported liking	3.39* (0.05)	−.25* (.05)	−.03 (.03)	3.38* (0.07)	.21* (.07)	−.23* (.04)	−0.01 (.09)	.46* (.09)	−.20* (.05)	0.75

Note. Standard errors are presented within parentheses.
* $p < .05$.

Promoted children showed an increase in emotional problems ($S2_P = .04$, with $SD = .02$; $d = .17$, $p < .05$) in the longer term; in contrast, retained children showed little change over the 4 years. However, the differences between the longer term slopes for the retained and promoted children did not attain statistical significance. Both retained and promoted children showed little change in conduct problems, peer-rated aggression, and peer-rated hyperactivity in the short term or longer term.

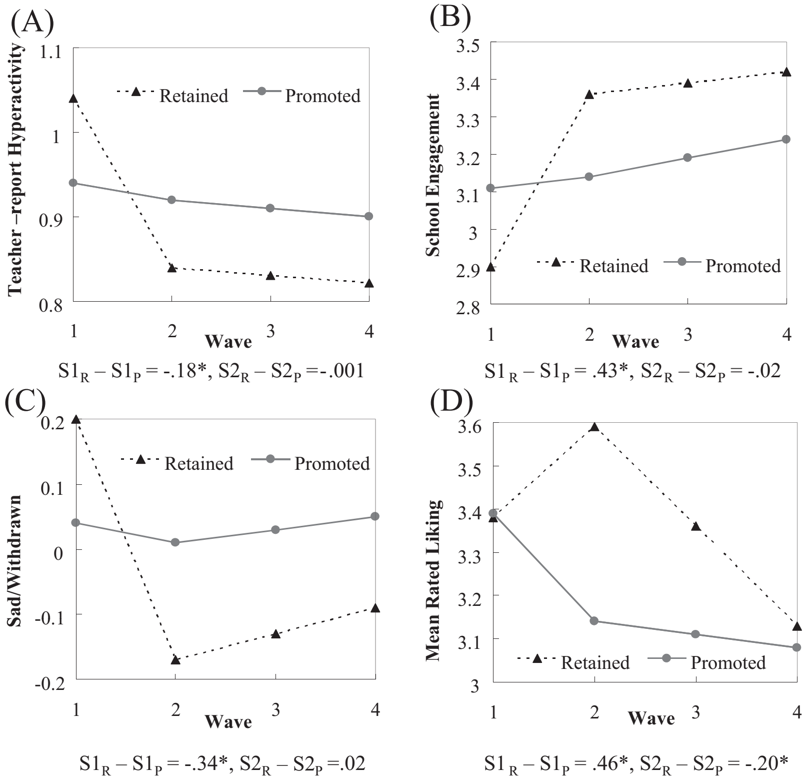
Model fit. Following Vonesh, Chinchilli, and Pu (1996), we used the conditional concordance correlation (CCC) to examine the fit of the piecewise model with retention as a Level 2 predictor for each of the teacher and peer report measures. Standard fit measures developed for use in confirmatory factor analysis strongly reflect the fit of the covariance structure, whereas the agreement between the observed and estimated individual responses is of primary interest in growth curve models (Wu, West, & Taylor, in press). Vonesh et al. developed the conditional CCC to assess the agreement between the observed and estimated individual responses in mixed-effects model settings. As a correlational measure, a value of 1 on the conditional CCC indicates a perfect fit and 0 indicates no agreement between the estimated and observed individual responses. As shown in Table 5, the piecewise model with retention as a Level 2 predictor results in high agreement between estimated and observed individual responses for all of the measures (conditional CCC ranged from .67 to .86), except for the teacher report for emotional problems, which was only moderate (conditional CCC = .31).

Regression Analyses on Student Report Measures

Table 7 shows the combined parameter estimates for the effect of grade retention on academic self-efficacy and school belonging at Grades 2 to 4. Grade retention had no significant effect on students' reports of academic self-efficacy at either Grade 2 or Grade 3 (.01, with $SE = .36$; 2.31, with $SE = 1.41$). However, retention did show a positive effect on students' reports of academic self-efficacy at Grade 4 (2.83, with $SE = 1.27$). Retained children reported having higher academic self-efficacy than promoted children at Grade 4, although the retained and promoted groups did not differ in academic self-efficacy in the shorter term. In contrast, grade retention had a positive effect on student reports of school belonging at Grade 2 (0.46, with $SE = .20$) but no substantial effect on school belonging at either Grade 3 or Grade 4 (0.40, with $SE = .26$; 0.42, with $SE = .31$). This pattern suggests that retained children had higher reports of school belonging than did promoted children during the repeat year but that this advantage disappeared in the longer term.

Discussion

Using two-piece linear growth curve models for teacher- and peer-reported outcomes, we found that retention in the first grade decreased teacher-rated hyperactivity and increased teacher-rated behavioral engagement of children in the repeat year. Retention also decreased peer-rated sad/withdrawn behaviors and increased peer liking in the repeat year. With the exception of peer liking, in the longer term, retained children kept the same growth rate in these outcomes as did the promoted children. These results indicate that retained children benefit from retention in terms of



(A) $S1_R - S1_P = -.18^*$, $S2_R - S2_P = -.001$

(B) $S1_R - S1_P = .43^*$, $S2_R - S2_P = -.02$

(C) $S1_R - S1_P = -.34^*$, $S2_R - S2_P = .02$

(D) $S1_R - S1_P = .46^*$, $S2_R - S2_P = -.20^*$

Figure 2. Estimated two-piece linear growth curves for (A) teacher-rated hyperactivity, (B) behavioral engagement, (C) sad/withdrawn, and (D) mean rated liking. The variables $S1_P$ and $S2_P$ represent the slopes in the short- and longer term for promoted children, and $S1_R$ and $S2_R$ represent the slopes in the short- and longer term for retained children. $^* p < .05$.

decreased teacher-rated hyperactivity, decreased peer-rated sad/withdrawn behaviors, and increased teacher-rated behavioral engagement in the short term and that this benefit was retained in the longer term. Conversely, the gain that retained children experienced in peer liking during the repeat of first grade was not maintained in the longer term. The results for several other psychosocial outcomes, including teacher-rated conduct problems and emotional problems as well as peer-rated hyperactivity and aggression, showed no short- or longer term effects of retention.

Table 7
Mean Differences Between the Retained and Promoted Groups on Academic Self-Efficacy and School Belonging at Waves 2, 3, and 4, Controlling for the Baseline Measure (Grade 1) and Probability of Being Retained at the End of Grade 1

	Effect of retention		
Self-report measure	Wave 2	Wave 3	Wave 4
Academic self-efficacy	0.01 (0.36)	2.31 (1.41)	2.83* (1.27)
School belonging	0.46* (0.20)	0.40 (0.26)	0.42 (0.31)

Note. Standard errors are presented within parentheses.
$^* p < .05$.

The regression results suggest that retention had a positive effect on children's sense of school belonging in the short term but that this benefit did not last. Retention had a positive effect on children's perceived academic competence at Year 4. The lack of a positive effect of retention on academic self-efficacy in Years 2 and 3 may be due to the lower association between children's perceived academic competence and objective indicators of academic achievement in Grades 1–2 compared with Grades 3 and 4 (Marsh, Craven, & Debus, 1991).

Our results also should be interpreted in the context of these children's academic performance across the same 4 years. In a companion article with this same longitudinal data set, Wu et al. (2008a) found a positive slope for retention on reading and math grade standard scores in the short term (Years 1–2) and a negative slope for retention in the longer term (Years 2–4). The short-term results for many of the psychosocial variables in the current study are similar to the short-term results for achievement. Consistent with the big fish, little pond effect (Marsh & Craven, 2002), the change in the frame of reference for retained and promoted students favored the retained child in the short term. However, once the child completed the repeat year, the results became more mixed. Results for teacher-rated hyperactivity, sad/withdrawn behaviors, and behavioral engagement showed sustained positive effects of retention. In contrast, both peer-rated acceptance and self-rated sense of school belonging among retained children relative to promoted children decreased substantially after the repeat year. By Year 4, retained students no longer differed from promoted students on either peer acceptance[6] or school belonging.

A lack of maturity is one of the most common reasons teachers give for decisions to retain a child (Tomchin & Impara, 1992). Our results suggest that retained students do improve on teacher ratings of two aspects of psychosocial adjustment that develop rapidly during this age period. Teacher-rated hyperactivity–inattentiveness (i.e., restless, easily distracted, acting without thinking first) and behavioral engagement (i.e., persisting on tasks, trying hard, making plans. and following through) both improved. Retained children were also viewed by peers as less withdrawn and sad, behaviors that teachers might describe as socially immature. Finally, at Year 4, the retained children reported higher academic self-efficacy. Thus, our results offer some support to teachers' judgments that the "gift of an extra year" benefits students, at least through Year 4. The finding of positive effects of retention on perceived academic self efficacy and internalizing behaviors is consistent with results of Hong and Yu (2008), a study that also used propensity scores to minimize selection effects. Together, these two studies suggest that retention in kindergarten and first grade do not harm students' psychosocial adjustment, at least through Grade 4.

The negative effects of retention found in this study are the worsening longer term trajectory for peer liking for retained students and the decrease in perceived school belonging. It may be that the negative longer term trajectory for grade-level achievement for retainees in Years 2–4 (Wu et al., 2008a) accounts for this negative effect. That is, children's success in meeting classroom academic expectations may affect their social standing (Gleason et al., 2007). Another possible explanation for these findings is that with increasing age, the label of "retainee" accrues negative, stigmatizing connotations. The evidence for such a developmental shift is indirect. When children in Grades 1 through 6 were asked to rate the stressfulness of numerous events, older children perceived being retained in grade as more stressful than did students in younger grades (Yamamoto & Byrnes, 1987). In Texas, the state in which the current study was conducted, third grade is the first year that students are required to pass a test of grade-level academic competencies in order to be promoted to the next grade. Thus, increased student awareness and apprehension about grade retention may contribute to the negative longer term effect of retention on peer acceptance and perceived school belonging. Additionally, over time, retained children, who are overage for their grade, may begin to disassociate from their grade peers and affiliate more with their same-age rather than same-grade peers, which could negatively impact their social acceptance in the classroom.

Strengths and Limitations of the Study

We matched retained and promoted children on the basis of their propensity scores estimated with a comprehensive set of background variables. This method substantially reduced selection bias and strengthened causal inference about the effect of grade retention in our nonrandomized study. Our analyses showed that we were able to achieve good balance between groups by matching. Additional analyses including less balanced background variables as covariates did not substantially alter the results. Retention also had no effect on the initial status (intercept) for all of the outcomes, which further supports the success of the matching. We maximized statistical power by using an optimal matching procedure that maximized our sample size and made the greatest possible use of our available data (Ming & Rosenbaum, 2000).

The primary limitations of the present study are associated with the collection of only four waves of data to date. The number of waves of data restricts the forms of growth that could be investigated. Limited to four waves of data collection, we could only fit a linear growth curve to represent the trajectory for short-term and long-term change. We were not able to examine possible nonlinear longer term growth trajectories (e.g., growth toward an asymptote) given that we had only three waves of postretention data. Consequently, we could not determine whether the trajectories we observed for some of our outcomes would continue or would attenuate over a longer period of time. In addition, the promoted children were only in fourth grade at the end of the study, before the transition to middle school and the beginning of adolescence. Many important changes in psychosocial outcome only arise at the point of school transitions, the change to a new developmental stage, or both (Coie et al., 1993; Roeser & Eccles, 1998; Wigfield & Eccles, 1994). We are currently collecting additional waves of data that will permit examination of possible nonlinear trajectories resulting from grade retention.

Because the psychosocial variables that were the focus of this investigation are associated with sociocultural factors, such as gender, language, socioeconomic status, and ethnicity, it is important to consider how such variables might interact with retention. However, a full investigation of such effects is outside the scope of the current investigation.

A final limitation in our study concerned missing data in the outcome variables. Although we carefully tracked all participants and used procedures that help maintain high rates of participation (see Ribisl et al., 1996), some observations were missing and a few participants were lost to the study (e.g., moved to unknown location). We used full information maximum likelihood procedures to address missing data; these procedures provide proper adjustment of the results for all measured variables (Schafer & Graham, 2002). However, if both the outcomes and missingness were related to unmeasured variables, then our estimates of the growth trajectories might have been biased, although they would nearly always be less biased than unadjusted estimates. No statistical test exists that allows us to test whether missingness is dependent on only measured variables (i.e., missing at random; Little & Rubin, 2002).

Summary and Conclusion

We found that retained children, relative to promoted children, benefited from retention in both the short and longer term with respect to decreased teacher-rated hyperactivity, peer-rated sadness and withdrawal, and increased teacher-rated behavioral engagement. Three years after retention, retained children reported higher academic competence than did matched promoted children. Our findings also point to possible trouble on the horizon. Particularly troubling are results that show a short-term increase in

[6] Difference = .05, with *SD* = .08.

peer-rated liking for the retained students, followed by a rapid decrease after the repeat year, as well as a short-term increase in school belonging that dissipated by Year 3. The companion study by Wu et al. (2008a) showed short-term improvement in math and reading achievement during the repeat year, followed by a rapid decline relative to grade mates as the children encountered new material. Future research is needed to test whether this "struggle–succeed–struggle" sequence for academic and social outcomes has long-term negative consequences for the academic motivation and achievement of retained children. These negative trajectories, if maintained, may portend some of the negative effects of retention in grade that have been observed in well-controlled studies of the effects of grade retention that have followed children into adolescence (e.g., Alexander et al., 2003; Pagani et al., 2001).

The abilities to pay attention, inhibit motor activity, persist on tasks, and manage one's emotions are aspects of self-regulation that develop rapidly during this age period (Blair, 2002; Kochanska et al., 1997). Retained students generally perform less well on measures of self-regulation (Jimerson et al., 1997; Willson & Hughes, 2009). The additional year of maturation affords retained students the opportunity to "catch up" to their (younger) grade mates in behavioral and emotional adjustment. Furthermore, improvements in adjustment are maintained through Year 4. Conversely, the longer term results for variables that may be less a function of maturation (i.e., peer liking and school belonging) are less encouraging. The label of "retainee" may have a stigmatizing effect on peers and contribute to the fading of the short-term benefit of retention on children's sense of belonging to school.

The finding of different effects across time and across domains of adaptation is consistent with transactional models of development (Lerner, 1998). Despite benefits through fourth grade, retention may create vulnerabilities that do not appear until the middle grades. For example, the struggle–succeed–struggle sequence in meeting grade-level academic competencies may contribute to the belief that one's academic outcomes are outside one's control, undermining academic motivation. Similarly, retention may bestow social advantages in the short term but have detrimental effects on social acceptance in the longer term, as students become more sensitive to being over-age for grade. Longitudinal studies that test such dynamic models of the effects of early grade retention hold promise for clarifying how retention affects long-term academic and social adaptation. Such an understanding is necessary to realize the benefits and avoid the costs of this educational intervention.

References

Ackerman, B. P., Izard, C. E., Kobak, R., Brown, E. D., & Smith, C. (2007). Relation between reading problems and internalizing behavior in school for preadolescent children from economically disadvantaged families. *Child Development, 78,* 581–596.

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., Entwisle, D. R., & Dauber, S. L. (1994). *On the success of failure.* Cambridge, United Kingdom: Cambridge University Press.

Alexander, K. A., Entwisle, D. R., & Dauber, S. L. (2003). *On the success of failure: A reassessment of the effects of retention in the primary grades.* Cambridge, United Kingdom: Cambridge University Press.

Anderson, G. E., Jimerson, S. R., & Whipple, A. D. (2005). Student ratings of stressful experiences at home and school: Loss of a parent and grade retention as superlative stressors. *Journal of Applied School Psychology, 21,* 1–20.

Appleton, J. J., Christenson, S. L., Kim, D., & Reschly, A. L. (2006). Measuring cognitive and psychological engagement: Validation of the student engagement instrument. *Journal of School Psychology, 44,* 427–445.

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

Beidel, D. C. (1991). Social phobia and overanxious disorder in school-age children. *Journal of the American Academy of Child & Adolescent Psychiatry, 30,* 545–552.

Bierman, K. L. (2004). *Understanding and treating peer rejection.* New York: Guilford Press.

Blair, C. (2002). School readiness: Integrating cognition and emotion in a neurobiological conceptualization of children's functioning at school entry. *American Psychologist, 57,* 111–127.

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

Brophy, J. (1983). Research on the self-fulfilling prophecy and teacher expectations. *Journal of Educational Psychology, 75,* 631–661.

Buhs, E. S. (2005). Peer rejection, negative peer treatment, and school adjustment: Self-concept and classroom engagement as mediating processes. *Journal of School Psychology, 43,* 407–424.

Buhs, E. S., & Ladd, G. W. (2001). Peer rejection as an antecedent of young children's school adjustment: An examination of mediating process. *Developmental Psychology, 37,* 550–560.

Buhs, E. S., Ladd, G. W., & Herald, S. L. (2006). Peer exclusion and victimization: Processes that mediate the relation between peer group rejection and children's classroom engagement and achievement? *Journal of Educational Psychology, 98,* 1–13.

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

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

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Coie, J. D., Watt, N. F., West, S. G., Hawkins, J. D., Asarnow, J. R., Markman, H. J., Ramey, S. L., Shure, M. B., & Long, B. (1993). The science of prevention: A conceptual framework and some directions for a national research program. *American Psychologist, 48,* 1013–1022.

Cole, D. A., Martin, J. M., & Powers, B. (1997). A competency-based model of child depression: A longitudinal study of peer, parent, teacher, and self-evaluations. *Journal of Child Psychology and Psychiatry, 38,* 505–514.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychological Methods, 6,* 330–351.

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.

Elliot, A. J., Sheldon, K. M., & Church, M. A. (1997). Avoidance personal goals and subjective well-being. *Personality and Social Psychology Bulletin, 23,* 915–927.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8,* 430–457.

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

Festinger, L. A. (1954). A theory of social comparison processes. *Human Relations, 7,* 117–140.

Finn, J. D. (1989). Withdrawing from school. *Review of Educational Research, 59,* 117–142.

Flook, L., Repetti, R. L., & Ullman, J. B. (2005). Classroom social experiences as predictors of academic performance. *Developmental Psychology, 41,* 319–327.

Fredericks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74,* 59–109.

Funder, D. C., & West (1993). (Eds.). Consensus, self–other agreement, and accuracy in personality judgment. *Journal of Personality, 61*(4) [Special issue].

Furrer, C., & Skinner, E. (2003). Sense of relatedness as a factor in children's academic engagement and performance. *Journal of Educational Psychology, 95,* 148–162.

Gleason, K. A., Kwok, O., & Hughes, J. N. (2007). The short-term effect of grade retention on peer relations and academic performance of at-risk first graders. *The Elementary School Journal, 107,* 327–340.

Goodenow, C. (1993). The psychological sense of school membership among adolescents: Scale development and educational correlates. *Psychology in the Schools, 30,* 79–90.

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.

Gootman, E. (2005, March 19). One in 3 city 4th graders may not advance to 5th. *The New York Times,* p. B5.

Gottfredson, D. C., Fink, C. M., & Graham, N. (1994). Grade retention and problem behavior. *American Educational Research Journal, 31,* 761–784.

Greenwood, C. R. (1991). Longitudinal analysis of time, engagement, and achievement in at-risk versus non-risk students. *Exceptional Children, 57,* 521–536.

Hagborg, W. J. (1998). An investigation of a brief measure of school membership. *Adolescence, 33,* 461–468.

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

Harter, S., & Pike, R. (1981). *The Pictorial Scale of Perceived Competence and Social Acceptance for Young Children.* Denver, CO: University of Denver.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58,* 47–77.

Herman, K. C., Lambert, S. F., Reinke, W. M., & Ialongo, N. S. (2008). Low academic competence in first grade as a risk factor for depressive cognitions and symptoms in middle school. *Journal of Counseling Psychology, 55,* 400–410.

Hill, C. R., & Hughes, J. N. (2007). An examination of the convergent and discriminant validity of the Strengths and Difficulties Questionnaire. *School Psychology Quarterly, 22,* 380–406.

Hinshaw, S. P. (1992). Externalizing behavior problems and academic underachievement in childhood and adolescence: Causal relationships and underlying mechanisms. *Psychological Bulletin, 111,* 127–155.

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.

Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis, 27,* 205–224.

Hong, G., & Yu, B. (2008). Effects of kindergarten retention on children's social–emotional development: An application of propensity score method to multivariate, multilevel data. *Developmental Psychology, 44,* 407–421.

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: Guilford Press.

Hughes, J. N., Luo, W., Kwok, O., & Loyd, L. (2008). Teacher–student support, effortful engagement, and achievement: A three-year longitudinal study. *Journal of Educational Psychology, 100,* 1–14.

Hughes, J. N., & Zhang, D. (2007). Effects of the structure of classmates' perceptions of peers' academic abilities on children's academic self-concept, peer acceptance, and classroom engagement. *Journal of Contemporary Educational Psychology, 32,* 400–419.

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

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

Jimerson, S. R. (2001b). A synthesis of grade retention research: Looking backward and moving forward. *The California School Psychologist, 6,* 47–59.

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.

Jussim, L. (1986). Self-fulfilling prophecies: A theoretical and integrative review. *Psychological Review, 93,* 429–445.

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.

Kochanska, G., Murray, K., & Coy, K. C. (1997). Inhibitory control as a contributor to conscience in childhood: From toddler to early school age. *Child Development, 68,* 263–277.

Ladd, G. W., Birch, S. H., & Buhs, E. S. (1999). Children's social and scholastic lives in kindergarten: Related spheres of influence? *Child Development, 70,* 1373–1400.

Ladd, G. W., Buhs, E. S., & Seid, M. (2000). Children's initial sentiments about kindergarten: Is school liking an antecedent of early classroom participation and achievement? *Merrill-Palmer Quarterly, 46,* 255–279.

Lerner, R. M. (1998). Theories of human development: Contemporary perspectives. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology* (Vol. 1, 5th ed., pp. 1–24). New York: Wiley.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

Lorence, J. (2006). Retention and academic achievement research revisited from a United States perspective. *International Education Journal, 7,* 731–777.

Mac Iver, D. (1988). Classroom environments and the stratification of pupils' ability perceptions. *Journal of Educational Psychology, 80,* 495–505.

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

Marsh, H. W., & Craven, R. (2002). The pivotal role of frames of reference in academic self-concept formation: The big fish little pond effect. In F. Pajares & T. Urdan (Eds.), *Adolescence and education* (Vol. 2, pp. 83–123). Greenwich, CT: Information Age.

Marsh, H. W., Craven, R. G., & Debus, R. (1991). Self-concepts of young children 5 to 8 years of age: Measurement and multidimensional structure. *Journal of Educational Psychology, 83,* 377–392.

Masten, A. S., Roisman, G. I., Long, J. D., Burt, K. B., Obradovic, J., Riley, J. R., et al. (2005). Developmental cascades: Linking academic achievement and externalizing and internalizing symptoms over 20 years. *Developmental Psychology, 41,* 733–746.

Maughan, B., Rowe, R., Loeber, R., & Stouthamer-Loeber, M. (2003). Reading problems and depressed mood. *Journal of Abnormal Child Psychology, 31,* 219–229.

McCombs Thomas, A., Armistead, L., Kempton, T., Lynch, S., Forehand, R., Nousianen, S., et al. (1992). Early retention: Are there long-term beneficial effects? *Psychology in the Schools, 29,* 342–347.

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

McDonald, A. S. (2001). The prevalence and effects of test anxiety in school children. *Educational Psychology, 21,* 89–101.

Meisels, S. J., & Liaw, F. R. (1993). Failure in grade: Do retained students catch up? *Journal of Educational Research, 87,* 69–77.

Miles, S. B., & Stipek, D. (2006). Contemporaneous and longitudinal associations between social behavior and literacy achievement in a sample of low-income elementary school children. *Child Development, 77,* 103–117.

Ming, K., & Rosenbaum, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics, 56,* 118–124.

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

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

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research, 66,* 543–578.

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

Pianta, R. C., Tietbohl, P. J., & Bennett, E. M. (1997). Differences in social adjustment and classroom behavior between children retained in kindergarten and groups of age and grade matched peers. *Early Education and Development, 8,* 137–152.

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.

Realmuto, G. M., August, G. J., Sieler, J. D., & Pessoa-Brandao, L. (1997). Peer assessment of social reputation in community samples of disruptive and nondisruptive children: Utility of the revised class play method. *Journal of Clinical Child Psychology, 26,* 67–76.

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

Ribisl, K. M., Walton, M. A., Mowbray, C. T., Luke, D. A., Davidson, W. A., & Bootsmiller, B. J. (1996). Minimizing participant attrition in panel studies through the use of effective retention and tracking strategies: Review and recommendations. *Evaluation and Program Planning, 19,* 1–25.

Risi, S., Gerhardstein, R., & Kistner, J. (2003). Children's classroom peer relationships and subsequent educational outcomes. *Journal of Clinical Child and Adolescent Psychology, 32,* 351–361.

Roderick, M., & Nagaoka, J. (2005). Retention under Chicago's high stakes testing program: Helpful, harmful, or harmless? *Educational Evaluation and Policy Analysis, 27,* 309–340.

Roeser, R. W., & Eccles, J. S. (1998). Adolescents' perceptions of middle school: Relation to longitudinal changes in academic and psychological adjustment. *Journal of Research on Adolescence, 8,* 123–158.

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.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association, 100,* 322–331.

Rubin, D. B. (2006). *Matched sampling for causal effects.* New York: Cambridge University Press.

Sameroff, A. J. (1975). Transactional models in early social relations. *Human Development, 18,* 65–79.

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

Schafer, J. L., & Kang, J. D. Y. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods, 13,* 279–313.

Schunk, D. H., & Zimmerman, B. J. (2006). *Competence and control beliefs: Distinguishing the means and ends.* Mahwah, NJ: Erlbaum.

Schwartz, D., Gorman, A. H., Duong, M. T., & Nakamoto, J. (2008). Peer relationships and academic achievement as interacting predictors of depressive symptoms during middle childhood. *Journal of Abnormal Psychology, 117,* 289–299.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton-Mifflin.

Shepard, L. A., Smith, M. L., & Marion, S. F. (1996). Failed evidence on grade retention. *Psychology in the Schools, 33,* 251–261.

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

Skinner, E. A., Zimmer-Gembeck, M. J., & Connell, J. P. (1998). Individual differences and the development of perceived control. *Monographs of the Society for Research in Child Development, 63*(2–3, Whole No. 204).

Stipek, D. J. (1981). Children's perceptions of their own and their classmates' ability. *Journal of Educational Psychology, 73,* 404–410.

Stipek, D. J., & Tannatt, L. M. (1984). Children's judgments of their own and their peers' academic competence. *Journal of Educational Psychology, 76,* 75–84.

Strauss, C. C., Lahey, B. B., & Jacobsen, R. H. (1982). The relationship of three measures of childhood depression to academic underachievement. *Journal of Applied Developmental Psychology, 3,* 375–380.

Texas Education Agency. (2005). *Grade-level retention in Texas public schools, 2003–04* (Document No. GE06 601 01). Austin, TX: Author.

Tomchin, E. M., & Impara, J. C. (1992). Unraveling teachers' beliefs about grade retention. *American Educational Research Journal, 29,* 199–223.

Trzesniewski, K. H., Moffitt, T. E., Caspi, A., Taylor, A., & Maughan, B. (2006). Revisiting the association between reading achievement and antisocial behavior: New evidence of an environmental explanation from a twin study. *Child Development, 77,* 72–88.

Valeski, T. N., & Stipek, D. J. (2001). Young children's feelings about school. *Child Development, 72,* 1198–1213.

Vonesh, E. F., Chinchilli, V. M., & Pu, K. (1996). Goodness-of-fit in generalized nonlinear mixed-effects models. *Biometrics, 52,* 572–587.

Weinstein, R. S., Marshall, H. H., Sharp, L., & Botkin, M. (1987). Pygmalion and the student: Age and classroom differences in children's awareness of teacher expectations. *Child Development, 58,* 1079–1093.

West, S. G., & Thoemmes, F. (2008). Equating groups. In P. Alasuutari, L. Bickman, & J. Brannen (Eds.), *The SAGE handbook of social research methods* (pp. 414–430). London: Sage.

West, S. G., & Thoemmes, F. (in press). Campbell's and Rubin's perspectives on causal inference. *Psychological Methods.*

Wigfield, A., & Eccles, J. S. (1994). Children's competence beliefs, achievement values, and general self-esteem: Change across elementary and middle school. *Journal of Early Adolescence, 14*(2), 107–138.

Wigfield, A., Eccles, J. S., Yoon, K. S., Harold, R. D., Arbreton, A. J. A., & Freedman-Doan, C., et al. (1997). Change in children's competence beliefs and subjective task values across the elementary school years: A 3-year study. *Journal of Educational Psychology, 89,* 451–469.

Willson, V. L., & Hughes, J. N. (2009). Who is retained in first grade? A psychosocial perspective. *The Elementary School Journal, 109,* 251–266.

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

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson III tests of achievement.* Itasca, IL: Riverside.

Wu, W., West, S. G., & Hughes, J. N. (2008a). Effect of retention in first grade on children's achievement trajectories over 4 years: A piecewise growth analysis using propensity score matching. *Journal of Educational Psychology, 100,* 727–740.

Wu, W., West, S. G., & Hughes, J. N. (2008b). Short-term effects of grade retention on growth rate of Woodcock–Johnson III broad math and reading scores. *Journal of School Psychology, 46,* 85–105.

Wu, W., West, S. G., & Taylor, A. B. (in press). Evaluating model fit for growth curve models: Integration of fit indices from SEM and MLM frameworks. *Psychological Methods, 14,* xxx–xxx.

Yamamoto, K., & Byrnes, D. (1987). Primary children's ratings of the stressfulness of experiences. *Journal of Research in Childhood Education, 2,* 117–121.

Zettergren, P. (2003). School adjustment in adolescence for previously rejected, average and popular children. *British Journal of Educational Psychology, 73,* 207–221.

Received February 8, 2008
Revision received April 20, 2009
Accepted May 18, 2009 ■