Routledge
Taylor & Francis Group

Journal of Research on Educational Effectiveness

ISSN: 1934-5747 (Print) 1934-5739 (Online) Journal homepage: http://www.tandfonline.com/loi/uree20

Rethinking Early Elementary Grade Retention: Examining Long-Term Academic and Psychosocial Outcomes

Sophia H. J. Hwang & Elise Cappella

To cite this article: Sophia H. J. Hwang & Elise Cappella (2018) Rethinking Early Elementary Grade Retention: Examining Long-Term Academic and Psychosocial Outcomes, Journal of Research on Educational Effectiveness, 11:4, 559-587, DOI: 10.1080/19345747.2018.1496500

To link to this article: https://doi.org/10.1080/19345747.2018.1496500

JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
2018, VOL. 11, NO. 4, 559–587
https://doi.org/10.1080/19345747.2018.1496500

INTERVENTION, EVALUATION, AND POLICY STUDIES

Check for updates

Rethinking Early Elementary Grade Retention: Examining Long-Term Academic and Psychosocial Outcomes

Sophia H. J. Hwang and Elise Cappella

ABSTRACT

Grade retention, the practice of requiring a student to remain in the same grade the following year, disproportionately affects students with sociodemographic risk and facing academic challenges. Each year, the United States spends $20 billion on retention and two million children are retained. Extant studies examining early elementary grade retention generally focus on short-term effects and academic outcomes; little is known about long-term effects on academic and psychosocial outcomes in the middle grades. The current study uses propensity score methods and a national data set to estimate the effect of first- or second-grade retention on academic achievement and psychosocial outcomes six or seven years later. By comparing students who were retained to students who were similar on observed characteristics but otherwise promoted, we generate causal estimates that show a statistically significant negative effect of retention on reading achievement. Significant and robust effects were not consistently detected for other academic or psychosocial outcomes. As grade retention is a widely used educational intervention, implications for its effectiveness from a policy and practice perspective are discussed.

KEYWORDS

grade retention
propensity score
adolescence
reading
psychosocial

Grade retention is a common and controversial educational practice in the United States. Retention—requiring a student who has underperformed academically to remain at the same grade level the following year (Jackson, 1975)—occurs for 2.4% of U.S. students across all grades (Warren, Hoffman, & Andrew, 2014). It is most commonly implemented in first grade, with retention rates for first-grade students at 6.2% nationally (Warren et al., 2014) and similar rates reported regionally (Cannon & Lipscomb, 2011; Karweit, 1999). Grade retention supporters claim it is effective as a remedial intervention giving students "the gift of time" to improve their academic achievement (Smith & Shepard, 1988; Tomchin & Impara, 1992). Scholars who advocate for social promotion (i.e., moving students to the next grade even if current performance standards are not met; Reschly & Christenson, 2013) argue that students would fare better academically if they were not retained.

For decades, researchers have examined the effects of early-grade retention on short-term academic outcomes, with early consensus classifying retention as an ineffective and

CONTACT Sophia H. J. Hwang sophia.hwang@nyu.edu 246 Greene Street, 8th Floor, New York, NY 10003, USA.
Department of Applied Psychology, New York University, New York, New York, USA.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uree

harmful remedial intervention (see meta-analyses: Holmes, 1989; Jimerson, 2001). The results of recent, methodologically rigorous studies vary, however, from no detected effects to significant negative effects to differential effects across various intervals of time and types of outcomes (e.g., Gleason, Kwok, & Hughes, 2007; Hong & Raudenbush, 2005, 2006; Hong & Yu, 2007, 2008; Im, Hughes, Kwok, Puckett, & Cerda, 2013). Fewer studies examine the effects of retention beyond elementary school into early adolescence on both academic and psychosocial outcomes critical to subsequent school and work success. To address this gap, the current study uses propensity score methods and a national longitudinal data set (Early Childhood Longitudinal Study, Kindergarten Class 1998–1999: ECLS-K; U.S. Department of Education, 2009) to estimate the effect of early-grade retention on academic achievement and psychosocial outcomes six or seven years later. The overall goal is to clarify whether retention is effective for students in first or second grade, and to inform education policy and practice regarding the potential costs and benefits of early-grade retention.

Influence of Early Elementary Retention on the Middle Grades: Context and Theory

Two million children are retained annually in the United States at a cost of approximately 20 billion dollars per year (Eide & Goldhaber, 2005). Over the past century, retention rates have fluctuated (Bali, Anagnostopoulos, & Roberts, 2005), with higher rates coinciding with more recent policy efforts to end social promotion (Clinton, 1998; Hursh, 2007). From a pedagogical standpoint, retention decreases classroom heterogeneity in achievement, which theoretically eases the instructional demands on teachers (Shepard & Smith, 1988). With limited access to alternative remediation strategies, policy directives such as No Child Left Behind (NCLB, 2002) may have indirectly led to a reliance on retention as an intervention for underperforming students (Lorence, 2009). For instance, administrators may have been compelled to retain low-performing students so schools could meet adequate yearly progress on high-stakes tests (Hursh, 2007).

Studies of retention indicate that its implementation disproportionately affects sociodemographically at-risk students facing academic challenges (Xia & Kirby, 2009). The exception is kindergarten retention, in which white students are overrepresented (Hong & Yu, 2007) and parents cite immaturity rather than academic difficulty as the primary reason for retention (Karweit, 1999). Beyond kindergarten, in contrast, retention disproportionately affects males, non-Hispanic Blacks, and students from lower income families (Warren et al., 2014; Xia & Kirby, 2009). Prior to retention, the students have lower academic abilities, social skills, and emotional adjustment than their promoted peers (Xia & Kirby, 2009). Given legal and ethical concerns around educational equity in schools, districts, and states, policies such as retention warrant rigorous attention by scholars and policymakers (Marsh, Gershwin, Kirby, & Xia, 2009).

Because retention is intended to impact student outcomes in the short- and long-term, developmental theory (Elder, 1994) helps explain how retention may (or may not) work. If retention were an effective remediation generating academic gains, it may interrupt a pattern of negative interactions for a struggling student and start a "positive cascade" (Masten & Cicchetti, 2010). Specifically, it may foster the student's sense of

self-efficacy (Bandura, 2006), perhaps due to the "big fish, little pond" effect (Marsh & Craven, 2002), and set the foundation for positive, distal competencies to arise in adolescence (Masten & Coatsworth, 1998). Teachers are in a position to reinforce these initial positive messages because in comparison to younger, less-experienced peers, the retained student appears to excel. These positive social comparisons (Festinger, 1954) may influence various interactions and settings throughout elementary school, and lead to lasting effects in the middle grades.

Alternatively, retention may leave the struggling student to continue a "negative cascade" (Masten & Cicchetti, 2010). In this case, the repetition of curricular content is ineffective and retention becomes a form of academic tracking (Alexander et al., 2003). As students continue to struggle in the same grade, they may develop a negative attitude toward and begin to disengage from schooling (Jimerson et al., 2002; Jimerson & Ferguson, 2007). In addition, as children progress through middle childhood, they become increasingly aware of stigma (Hong & Yu, 2007), and may sense that their peers view them to be "bad" or "stupid" (labeling theory; Becker, 1963). Teachers also may have lowered expectations for retained students (i.e., expectancy theory; Weinstein, 2002), which may fuel the students' own negative self-perceptions.

The dynamic influence of retention can "snowball" over time and affect adolescent outcomes. However, these change processes do not have to be stable and linear. Short-term retention effects may not necessarily predict long-term effects (Adolph & Robinson, 2008). *Dynamic interactionism,* a tenet of developmental systems theory, underscores how individual–environment interactions are changing, reciprocal, and bidirectional (Gottlieb, 1992). To illustrate this principle, a student once on a positive cascade may not remain on that trajectory, or some retainees may "struggle-succeed-struggle" during their schooling (Gleason et al., 2007). Further, while it is possible for retention to have a varied and profound effect over time (Lerner, 2006), it can also have a differential influence depending on the outcome (academic or psychosocial). Broad-ranging psychosocial constructs, such as social self-concept (awareness of social support), internalizing behaviors (inward-focused concerns), locus of control (belief in one's influence), and self-esteem (self-worth) are interrelated, but distinct (Bong & Skaalvik, 2003). Thus, it is possible for a complex intervention like retention to have unique impacts.

This paper specifically focuses on outcomes in early adolescence, which is a challenging and turbulent time characterized by significant biological, psychosocial, and academic transitions (Steinberg, 2008). During this developmental period, peers increase in influence (Prinstein & Dodge, 2008) and feelings of belonging and social competence become salient (Norwalk, Hamm, Farmer, & Barnes, 2016). Retainees could have an advantage as social maturity and experience may help them navigate these transitions. However, positive middle school findings for retained students are scarce in prior empirical work. It may be that being overage, which is the case for most retained students, becomes prominent in the middle grades as they enter puberty earlier than their same-grade peers (Wu et al., 2010). A sample of sixth graders ranked retention to be their largest life stressor (Anderson, Jimerson, & Whipple, 2005), highlighting increased experiences of stigma and peer judgment. Thus, while social comparisons immediately after retention may be beneficial in elementary school, these comparisons in the middle grades could be detrimental, as retainees have enhanced sensitivity to perceptions by others.

Adolescence must also be contextualized by the additional challenges of navigating school transitions and changing relationships with teachers and peers (Roeser & Eccles, 1998; Wigfield & Eccles, 1994). Nearly 90% of students in the ECLS-K have a school transition in sixth or seventh grade (Cappella, Schwartz, Hill, Kim, & Seidman, 2017). Overall, students experience a decrease in achievement and motivation during the middle grades transition (Eccles et al., 1993). This may be explained by the mismatch between the students' academic, social, and developmental needs and the quality of the school environment (Eccles & Midgley, 1989). In contrast to elementary school, middle schools are generally larger and have more complex structures, with classroom-level ability grouping (e.g., block schedules, honors classes; Slavin, 1990) and varied curricula (Hoffer, 1992; Mulkey, Catsambis, Steelman, & Crain, 2005). Because middle schools are more formal and less personal, and students are taught by various teachers throughout the day, it is difficult for adolescents to form connections (Wigfield & Eccles, 1994). Although not all research teams concur (e.g., Im et al., 2013), some suggest that these middle school challenges may be exacerbated for retained students, as they are more likely than their promoted peers to experience difficulties after leaving elementary school (e.g., Alexander et al., 2003). Prior research has shown that attendance, test scores, grades, and behavior in the middle grades strongly predict high school achievement (Balfanz, 2009) and on-time graduation (Balfanz, Herzog, & Mac Iver, 2007). Thus, it is important to understand whether early elementary retention has a lasting detectable and significant negative effect over and above the baseline challenges of adolescence and being in the middle grades.

Methodological Considerations in Retention Research

Elementary retention has received substantial attention in education research, yet concerns about methodological rigor beset the early empirical work. A large body of early studies characterized retention as ineffective, with two pivotal meta-analyses (Holmes, 1989; Jimerson, 2001) prompting the National Association of School Psychologists to declare retention a "failed intervention" (NASP, 2011). Other scholars questioned the design and analytic methods of early studies (Allen, Chen, Willson, & Hughes, 2009; Lorence, 2006, 2009) and challenged their conclusions (see Alexander, Entwisle, & Dauber, 2003). For example, only five of the 63 studies in Holmes's (1989) meta-analysis adjusted for prior performance before retention (Lorence, 2009); only four studies in Jimerson's (2001) meta-analysis had adequate comparison groups and statistical controls (Lorence, 2006). These gaps make it difficult to discern whether outcomes for retained children were truly negative or the product of selection bias.

Mixed findings from more recent research may be partially attributed to differences in study methods (Allen et al., 2009; Lorence, 2006, 2009). Two common methodological approaches in retention studies—selection of comparison students with similar characteristics and inclusion of covariates in regression analysis—fail to fully account for potential pretreatment differences between retained and promoted students (Lorence, 2009). As noted by Allen et al. (2009), more rigorous studies report slightly positive effects (mean effect size: .04) and less rigorous ones report larger and more negative effects (mean effect size: −.30).

Mixed findings may also relate to whether a study utilizes a *same-age* or *same-grade* comparison in estimation of retention's effect. Same-age comparisons compare retained students at the same point in time to similarly aged peers who were socially promoted (e.g., Hong & Yu, 2007, 2008). Same-age comparisons better account for age-based developmental and social outcomes, and generally display more negative or null findings (Allen et al., 2009). In contrast, same-grade comparisons either compare retained students to their new, younger same-grade classmates during the repeating year (Alexander et al., 2003) or more commonly, compare retained students in the repeating year to the performance of their promoted peers when they were in the same grade (i.e., from the year prior, requiring data from adjacent years; Mariano & Martorell, 2013). Same-grade comparisons are more likely to show a positive effect of retention that fades over time (Allen et al., 2009), but a limitation is that estimates include changes due to maturation. Ultimately, the appropriateness of selecting a same-age or same-grade comparison depends on the research question. For instance, if researchers aim to illuminate how retained students would have fared if they were promoted, then a same-age comparison group is the appropriate counterfactual (i.e., this study). Conversely, if quantifying the effect of the repeating year on later outcomes is of interest, then a same-grade comparison is fitting. Both approaches are valuable as they highlight different aspects of retention's effects and consequences.

Last, the timing of the assessment of post-retention outcomes will influence the results. A concrete illustration is the regression discontinuity studies (Jacob & Lefgren, 2004; Roderick & Nagaoka, 2005) using same-grade comparisons that examined the 1996 discontinuation of social promotion in Chicago Public Schools. A short-term academic "spike" was measured in students repeating the third grade; these gains vanished two years later. Examination of Florida's third-grade test-based promotion policy found greater reading proficiency in retained students two years after retention using a same-grade comparison (Greene & Winters, 2007); this effect faded over time, but was still positive and significant when retained students were in the seventh grade (Winters & Greene, 2012). These examples underscore the importance of critically evaluating the quality, timing, and context of empirical evidence (Lorence, 2006; 2009).

Academic and Psychosocial Outcomes After Early-Grade Retention

On average, empirical studies using causal methods and appropriate comparison groups report negative or null effects of post-kindergarten early-grade retention on academic achievement (i.e., math and reading test scores). In a national longitudinal data set, Hong and Yu (2007) found that first-grade retention yielded negative effects on reading and math test scores one year later; these effects persisted over three years. In a prospective study in Texas, Wu, West, and Hughes (2008) compared 103 students retained in first grade to their one-to-one optimally matched propensity score pairs. First-grade retention did not have an effect on reading skills; however, it had a negative effect on the three-year growth rate of math skills. In a follow-up study using latent class trajectory analysis, a subset of retained students in the Texas sample that scored lower on academic assessments prior to retention displayed faster growth in reading and math achievement than initially peers who were also retained (Chen, Hughes, & Kwok, 2014).

However, on average, while there was an improvement immediately after retention, there was no lasting benefit by fifth grade (Moser, West, & Hughes, 2012). The authors note that students may actually be engaged in a "struggle-succeed-struggle" sequence that failed to be captured in the time frame of the study (Gleason et al., 2007).

Prior research using same-grade comparisons and focusing on test-based promotion or other state-specific retention policies suggests that retention may be effective (e.g., Chicago, Florida). In Texas, third-grade retainees experienced positive reading outcomes in comparison to socially promoted peers through the tenth grade (Lorence, 2014). In New York, positive short-term reading and math effects were found in seventh grade after fifth-grade retention (Mariano & Martorell, 2013). Although these quasi-experimental studies increase our understanding of retention as an intervention, generalizability is limited as they focus on high-stakes testing and target specific regions (Jacob & Lefgren, 2004; Greene & Winters, 2007; Winters & Greene, 2012).

Fewer studies have assessed the effect of retention on *non-test-score academic outcomes* (e.g., self-, peer-, or teacher-rated competencies), and the evidence is mixed. Hong and Yu (2008) examined outcomes one, two, and four years after kindergarten retention for 471 students. Retained students had increased competence and interest in academic learning two years (statistically significant) and four years (trend-level) later in comparison to same-age promoted peers, as rated by students, teachers, and parents. Gleason et al. (2007) compared first-grade retainees to socially promoted peers; in the repeating year, promoted students had higher academic engagement, and one year later, retainees had greater teacher- and peer-rated academic competence in comparison to same-age promoted peers.

Similarly, only a handful of empirical studies move beyond academic outcomes to examine the effects of first- or second-grade (rather than kindergarten) retention on *psychosocial outcomes*. Theoretically, retention could have positive effects on psychosocial indicators such as self-concept, academic efficacy, peer acceptance, or school belonging if retained students engage in social comparison with their new, younger classmates and experience a subsequent boost in confidence (Marsh & Craven, 2002). In the Texas study, those retained in first grade had higher peer acceptance in the treatment year in comparison to promoted peers; this was mediated by teacher- and peer-reported academic competence (Gleason et al., 2007). However, in the same sample, Wu et al. (2010) examined psychosocial outcomes one and three years after retention; retention decreased hyperactivity and increased behavioral engagement in the first and third year, but short-term benefits in school belonging and perceived peer liking did not endure.

Results between retention and other psychosocial outcomes such as internalizing behaviors, self-esteem, and locus of control have been similarly inconsistent. Retained students have been reported to have more internalizing behaviors (Pagani et al., 2001), fewer internalizing problems (Hong & Yu, 2008; Im et al., 2013), or no differences (McCombs-Thomas et al., 1992; Wu et al., 2010) in comparison to promoted peers. Alexander et al. (2003) found an increase in self-esteem during the repeating year that diminished over time, and lower levels of locus of control for first-grade retainees. Jimerson's (2001) meta-analysis included studies with negative, null, and positive outcomes for aggression, self-esteem, and locus of control for retained students. Rarely has

an empirical study examined multiple psychosocial outcomes, although it is plausible that retention has differential impacts when comparing across various outcomes in the middle grades. For example, after being retained, an early adolescent may engage in classroom misbehaviors, but depending on the school context, that student may be perceived by peers to be cool. This may lead to a high sense of social self-concept, but also internalizing behaviors if the student is concerned about academic performance (e.g., Masten & Coatsworth, 1998). However, psychosocial constructs remain understudied in the retention literature; thus, there is an opportunity to address this gap and examine their interrelations.

Limitations of Prior Work

Much of the extant literature focuses on academic rather than psychosocial outcomes, and has examined short-term (e.g., one to three years) rather than long-term (e.g., five to seven years) effects of retention on student development. There are several reasons why this is problematic. First, researchers have determined that social and emotional skills are important in their own right and for concurrent and subsequent well-being and achievement (Anderson, Jacobs, Schramm, & Splittgerber, 2000; Becker & Luthar, 2002; Farrington et al., 2012; Heckman, 2000). Second, some psychosocial and academic shifts may be evident in the short term; others may have a delayed presentation. These "sleeper effects" (Alexander et al., 2003; Gleason et al., 2007; Pagani et al., 2001) are not well understood and necessitate longer term developmental research. Third, if short-term positive gains after retention do not last into the middle grades and beyond, there may not be compelling evidence to support this costly and time-intensive intervention (Allen et al., 2009).

Only one known study has used causal methods to evaluate elementary school grade retention's effect on sixth-, seventh-, and eighth-grade academic and psychosocial outcomes. Im et al. (2013) examined 75 students in Texas who experienced retention at any point in elementary school (kindergarten to fifth grade) and followed them to eighth grade. The investigators hypothesized that during the transition into middle school, reading and math achievement, sense of school belonging, and behavioral engagement would be lower for retained students in comparison to promoted peers. However, using propensity score matching and piecewise growth modeling, they found no differences in trajectories between these groups; power analyses showed that they could adequately detect a meaningful effect, but a sample of 75 treatment students is small in comparison to other retention studies. Further, given the number of covariates (i.e., 67) for a small treatment group, model overfitting is a limitation that challenges the interpretation of the findings (similar limitations in prior work: e.g., Hughes, Cao, West, Smith, & Cerda, 2017; Wu et al., 2010). Although this investigation highlighted retention's impact during this crucial developmental period, research on a larger, national sample of students is needed to further corroborate or challenge these findings.

The current empirical literature examining grade retention has two additional gaps. First, most studies evaluate a specific region or state retention policy. Several published studies examining Chicago's ban on social promotion (Jacob & Lefgren, 2004, 2009; Roderick & Nagaoka, 2005), Florida's third-grade reading requirement for social promotion (Greene & Winters, 2007; Özek, 2015; Winters & Greene, 2012), and the studies

from a prospective Texas data set (including Im et al., 2013) provide valuable information; however, they have limited generalizability because they focus on a specific region. Hong and colleagues demonstrate a unique strength as they evaluate the national practice of retention rather than a specific regional policy—the current study extends this work. The second gap in the extant literature is the lack of emphasis on first- or second-grade retention. Many studies evaluated the impact of kindergarten retention (e.g., Hong & Raudenbush, 2006; Huang, 2014) but given that kindergarten retainees are likely to be held back for reasons of immaturity or "red-shirting," as well as for academic concerns, the academic implications of kindergarten retention may be less pronounced. Hong and Yu (2007) found negative academic effects for kindergarten retainees in the repeating year that faded out four years later, while first-grade retainees had persistent negative effects one and three years after retention. Several studies examine third-grade retention or later (e.g., sixth and eighth grade), as these grades are commonly associated with high-stakes testing and other accountability standards. However, only a few studies examine first- or second-grade retention, which is when retention is most likely to occur (e.g., Gleason et al., 2007; Wu et al., 2008).

Current Study

The present study builds upon prior work through the causal analysis of a nationally representative data set to investigate both academic achievement and psychosocial outcomes in early adolescence following early elementary grade retention. Using propensity score methodological approaches (Rosenbaum, 2002b), this study constructs a logical counterfactual group to compare retained students to other same-age students who would have been retained based on similar observed characteristics, but were promoted. A rich set of demographic, assessment, and school characteristic variables were included in the selection model. By estimating the effect of first- or second-grade retention on academic achievement and psychosocial outcomes six or seven years after retention, we can determine if retention is an effective intervention for early elementary school students. Although the treatment of retention may be different for these two grades, we believe they are likely to be similar, and due to data limitations in the ECLS-K, we cannot disentangle the retention episode (see Method section for details). In the absence of a sufficient body of literature examining first- or second-grade retention on longer term academic and socioemotional outcomes in a national sample, specific hypotheses cannot be generated. Thus, the current study's research questions are exploratory in order to investigate the presence, direction, and strength of the effects of grade retention.

Method

The current study involves the secondary analysis of the Early Childhood Longitudinal Study, Kindergarten Class 1998–1999 (ECLS-K; U.S. Department of Education, 2009), a nationally representative sample following a longitudinal cohort of 21,260 kindergarteners until the eighth grade. Data were collected by the National Center for Education Statistics (NCES; Tourangeau, Nord, Lê, Pollack & Atkins-Burnett, 2006) over seven waves: fall and spring of kindergarten (1998–1999), fall and spring of first grade

Table 1. Pretreatment demographic characteristics for retained and promoted students ($N = 5{,}586$).

	Retained ($n = 295$)		Promoted ($n = 5{,}291$)	
	n	%	*n*	%
Gender				
Male	170	57.63	2,551	48.21
Female	125	42.37	2,740	51.79
Race				
White	113	38.31	3,202	60.52
Black	81	27.46	540	10.21
Hispanic	72	24.41	939	17.75
Asian	8	2.71	310	5.86
Native Hawaiian or American Indian	14	4.75	168	3.18
Multiracial	7	2.37	128	2.42
Student has a disability at start of kindergarten	61	22.34	641	12.79
Region				
Northeast	50	16.95	1,003	18.96
Midwest	53	17.97	1,434	27.10
South	140	47.46	1,666	31.49
West	50	16.95	1,185	22.40
Urbanicity				
Urban	122	41.36	1,710	32.32
Suburban	80	27.12	2,098	39.65
Rural	91	30.85	1,442	27.25
	Mean	*SD*	Mean	*SD*
Spring 2000 reading IRT test score	−0.47	0.42	0.19	0.38
Spring 2000 math IRT test score	−0.48	0.41	0.12	0.37

(1999–2000), spring of third grade (2002), spring of fifth grade (2004), and spring of eighth grade (2007). Data utilized in the analysis were derived from the fall and spring of kindergarten (wave 1 and 2), spring of first grade (wave 4), and spring of eighth grade (wave 7).[1] The data are from multiple sources including administrator surveys, teacher assessments, direct child assessments, child questionnaires, and parent surveys.

Sample

According to First Findings from the Final Round of the ECLS-K (Walston, Rathbun & Germino Hausken, 2008), 13 percent of the students included in the 2007 data collection were enrolled in a grade below eighth grade (i.e., retained at least once after kindergarten in fall 1998). This study's sample is drawn from 5,586 students who meet the following inclusion criteria: (a) participated in all six data collection waves of focus[2]; (b) attended a public school throughout the study; (c) were first-time kindergarteners at the study's start; (d) were enrolled in first grade in spring 2000 (i.e., not retained in kindergarten after the study's first year or promoted ahead of time); and (e) have valid grade-level data in spring 2002 (i.e., students in ungraded classrooms or with missing/not ascertained grade-level data were excluded).

[1]Data were collected from only a subsample of students in the fall of first grade (wave 3), so that time point was not included in this analysis.

[2]ECLS-K longitudinal sampling weights were used to identify children who participated in the six data-collection waves of interest. In accordance with earlier ECLS-K studies (e.g., Reardon, Cheadle, & Robinson, 2009), this criterion was used so that sample attrition does not affect our treatment estimates and ensures that analyses were conducted on the same sample of eligible students.

Within this sample, 295 students who were retained once in either the first or second grade comprise the treatment group. These students were identified through teacher-report because they were recorded as being in the second grade in spring 2002, when the majority of the sample is in the third grade. As data were not collected each school year, we cannot determine in which year (i.e., first versus second grade) the student was retained. All remaining students in the sample who were reported as being in the third grade in spring of 2002 create the eligible pool of students ($n = 5{,}291$) from which we derive the comparison group for analysis (i.e., prematch comparison group of promoted students).

Table 1 presents the pretreatment demographic characteristics for retained and promoted students. The retained group consists of a higher proportion of males (58% versus 48%) and lower proportion of white students (38% versus 61%) as compared to the promoted group. Additionally, 22% of retained students were classified as having a disability (a proxy for special education), while 13% of promoted students had this status. Spring 2000 (first-grade) reading and math test scores were collected pretreatment. Prior to retention, students in the treatment group were performing at a much lower level than their peers who were promoted.

Measures

Independent Variable

In the current study, the treatment equals one instance of grade retention in either the first or the second grade (i.e., during the 1999–2000 or 2000–2001 academic year).[3] Continuously promoted children were in the third grade in spring 2002; those who were retained were one grade level behind.

Dependent Variables

The multidimensional academic and psychosocial outcomes of interest were collected in spring 2007 for all students, regardless of the student's grade level, yielding a same-age comparison. Academic outcomes include direct assessment of reading and math achievement, student self-report of reading and math competence, and teacher-report of reading competence.[4] Psychosocial outcomes include student self-report on social self-concept, internalizing behaviors, self-esteem, and locus of control.

Academic Outcomes. *Math* and *reading achievement* were measured by the direct assessment of students' math and reading skills on a 50- to 70-item two-stage test measuring latent ability in each subject based on the pattern of correct, incorrect, and

[3]Students retained in both first and second grade were excluded. Additionally, this study does not control for posttreatment retention episodes (i.e., occurring during or after third grade) because that would alter the outcome estimates (Gelman & Hill, 2007). Any student retained in either first or second grade, regardless of whether they were retained again later on, was part of the treatment group.

[4]Teacher-report of math competence was not included as an outcome because it was only collected for half of the sample in spring 2007; the other half of the students had teacher-report of science competence. Only English teachers were assigned to complete this report for all students.

omitted responses (Tourangeau, Nord, Lê, Sorongon, & Najarian, 2009). Scores are based on item response theory (IRT; Hambleton, Swaminathan & Rogers, 1991) and calibrated to be on the same scale. Specifically, this study uses the IRT-derived theta scores,[5] which are comparable across time and allow for the examination of growth over time; the theta score distribution range across all data-collection points is approximately −3 to 3 (Najarian, Pollack, & Sorongon, 2009). Theta reliability estimates are .87 (reading) and .92 (math) in spring 2007 (Najarian et al., 2009).

We examine student self-reported *reading competence* and *math competence* from the Self-Description Questionnaire (alpha = .76 and .89, respectively; Najarian et al., 2009). Each domain comprises four items on a four-point response scale (1 = *not at all true* to 4 = *very true*). Sample items include "I like reading" and "Math is one of my best subjects." We also use teacher-report of reading competence from the Academic Rating Scale to complement information gathered from the direct assessment (Tourangeau et al., 2009). The English teacher rated the student's oral (three items) and written expression skills (five items) on a five-point scale (1 = *poor* to 5 = *outstanding*). Reliability statistics for both reading competence measures were high (0.93 for oral competence and 0.96 for writing competence). Example items include "expresses analytical or critical thinking" and "employs English grammar and usage."

Psychosocial Outcomes. *Social self-concept* is a composite construct from the factor analysis of items related to social adjustment (Kim, Schwartz, Cappella, & Seidman, 2014). This measure includes five items rated on a five-point scale (1 = *never* to 5 = *always*) related to students' perception of peer support and acceptance (alpha = .89; e.g., "classmates care about me" or "classmates like me as I am").

In spring 2007, students self-reported their *internalizing problem behaviors* (e.g., feeling lonely, frustrated, and worrying about school) for eight items along a four-point scale (1 = *not at all true* to 4 = *very true*; alpha = .75). *Self-esteem* (alpha = .81) consists of seven items derived from the Rosenberg Self-Esteem Scale (RSE, Rosenberg 1965) such as "I feel good about myself" and "I feel I am a person of worth, the equal of other people." *Locus of control* (alpha = .68) consists of six items, for example, "I don't have enough control over the direction my life is taking" and "Chance and luck are very important for what happens in my life" (Najarian et al., 2009). For both self-esteem and locus of control, items were rated on a four-point scale (1 = *strongly agree* to 4 = *strongly disagree*); analyses utilized the scale score, which is the average of the standardized items with mean of zero and standard deviation of one.

Covariates

Informed by previous empirical work, 43 covariates were included in the propensity score model to adjust for the fact that potential confounders could predict both the likelihood of being retained and outcomes post-retention (see Appendix). Inclusion of these confounding covariates addresses whether there may be differential selection into treatment (retained) and control (not retained) groups. Covariates are drawn from data

[5]The corrected theta scores released by NCES in March 2010 were used for this study.

collected *during or before* spring of first grade, prior to the treatment of grade retention (i.e., from fall 1998, spring 1999, or spring 2000). Selection of covariates was based on the strength of the theorized relation to the treatment and outcome variables; those more strongly related to the outcome were prioritized (Gelman & Hill, 2007). Covariates include pretreatment measures of outcomes (when available), and additional administrative (child- and school-level), demographic, teacher-report, parent-report, and child assessment variables. Prior literature and theory (e.g., Willson & Hughes, 2009) have shown these to be important variables to consider.

Missing Information

The mean level of missingness for the data set's variables was 7.14%, ranging from 0–40%. To preserve all students meeting our inclusion criteria, we conducted an Imputation by Chained Equations (ICE; Little & Rubin, 2002) in STATA for missing values. Over 80 variables from both pretreatment and posttreatment time points (including Appendix covariates) were included in the imputation procedure. The single imputation model used additional variables as predictors and was specified for each type of variable (binary, categorical, or continuous); the ICE procedure is flexible as the model allows for different types of distributions (Horton & Kleinman, 2007). Ordered categorical variables with five or more categories were treated as continuous. Imputations were conducted stochastically to accommodate existing variation in data set. Visual inspection of the complete data set ensured that imputations had expected ranges and distributions. All analyses were conducted on the imputed data set.

Analytic Plan

This study has one main propensity score analytic approach (one-to-one nearest neighbor without replacement), on which an adjustment for multiple comparisons and a sensitivity analysis are employed. Caliper matching with replacement is a secondary propensity score approach that serves as a robustness check. Only findings that are: (a) significant in the main analysis ($p < .05$), (b) meaningful after correcting for multiple inference ($q < .10$ in the main analysis due to the conservative adjustment; e.g., Anderson, 2008); and (c) significant in the alternative specification ($p < .05$ in the caliper approach), are considered robust and consistent.

Regarding the retained students as the sample of interest, we turn to quasi-experimental methods to estimate the *average treatment effect on the treated* (ATT; Guo & Fraser, 2010). This enables our understanding of how the retained students would have performed on middle grade academic and psychosocial outcomes had they been promoted. This is expressed using the following formula:

$$\mathrm{E}[\mathrm{Y}(1) - \mathrm{Y}(0) \mid \mathrm{Z} = 1] = \mathrm{E}[\mathrm{Y}(1) \mid \mathrm{Z} = 1] - \mathrm{E}[\mathrm{Y}(0) \mid \mathrm{Z} = 1] \quad (1)$$

In Equation 1, E[Y(1)] and E[Y(0)] denote the expected marginal outcomes given that the students receive the treatment of retention (Z = 1). Because Y(1) and Y(0), the potential outcomes of being retained or not retained, respectively, cannot be observed for the same population of students, we employ propensity score matching. This procedure constructs logical counterfactual groups and compares treatment students only to

promoted students who would have been retained, based on similar observed characteristics (Hill & Reiter, 2006; Hill, Weiss & Zhai, 2011). Assuming the 43 child- and school-level covariates (see the appendix) capture all meaningful differences measured before retention and address the ignorability assumption (Rubin, 1978), we can estimate the effect of grade retention disentangled from other variables. Consistent with the stable unit value assumption (SUTVA; Rubin, 1978), we assume the retention of one student does not affect the potential outcome of another student, and that while retention may vary region to region and policy by policy, the treatment is essentially comparable for all units as it captures the practice of repeating the same grade level.

The methods used in this study are aligned with prior empirical work; each student in the analytic sample receives a propensity score estimated using a logistic regression, which summarizes the likelihood of being retained based on all pre-retention covariates (Rosenbaum, 2002b; Rosenbaum & Rubin, 1983). This propensity score acts as a one-number "scalar summary." We use two different propensity score methodological approaches to estimate the average treatment effect on the treated (see Stuart et al., 2009 for more information regarding propensity score approaches).

First, for the main analysis, we conduct a nearest neighbor (one-to-one) matching without replacement, allowing each treatment student to have one unique comparison student; this exploits the large number of students in the control group and generates treatment and control samples equal in size. Since ten outcomes were tested, a false discovery rate (FDR; Benjamini & Hochberg, 1995) correction was employed to address concerns regarding multiple comparisons (Benjamini, Krieger, and Yekutieli, 2006) for the main analysis. The FDR q-value adjusts for multiple inference and can be interpreted in the same way as a p-value (e.g., Anderson, 2008). For significant "naïve" p-values in the main analyses, adjusted q-value are reported to ensure confidence in the main analysis findings. Further, we conducted a sensitivity analysis using the Rosenbaum bounds approach (i.e., Wilcoxon signed rank test for matched pairs; DiPrete & Gangl, 2004; Rosenbaum, 2002b) on the main analysis to determine the strength of the effect of an unobserved covariate, related to both the treatment and outcome, that would increase the odds of being retained and alter the outcome's significance (Liu, Kuramoto, & Stuart, 2013).

The second propensity score approach we use is caliper matching with replacement, which serves as a robustness check. This approach implements stricter match criteria than the one-to-one matching, as set parameters determine how close the propensity scores of the control and treatment students must be in order to be considered a match (Guo & Fraser, 2010). Likely the caliper approach minimizes bias and yields more precise matches (see Results section for analytic details). Findings are considered robust if the results from the more restrictive caliper approach corroborate the significant outcomes in the main analysis.

We estimate retention's effect on the matched samples by conducting linear regressions (Equation 2) for each outcome of interest.

$$Y_i = \beta_0 + \tau Z_i + \Sigma \beta_c X_{ci} + \varepsilon_i \tag{2}$$

The outcome of interest is represented as Y_i for student i. The intercept is represented by β_0; τ is the treatment effect; and Z is the treatment assignment for student i ($0 =$ promoted, $1 =$ retained). Additionally, $\Sigma B_c X_{ci}$ captures student i's covariate

adjustment. This includes the 20 higher priority covariates listed in the appendix related to the likelihood of being retained and/or outcomes post-retention and variables that do not demonstrate sound balance in the propensity score matching procedure to reduce bias and generate more precise estimates (Hill, 2008; Rosenbaum 2002a; Rubin & Thomas, 2000). Lastly, ε_i denotes the error term.

Results

Main Analysis: One-to-One Nearest Neighbor Without Replacement

Diagnostics

Treatment students with propensity scores in the region of overlap with control group propensity scores were included in the analysis (i.e., "common support"; Caliendo & Kopeinig, 2008; Heckman, LaLonde, & Smith, 1999). Only eight students of the total 295 who were retained were not included because they did not meet the condition of sufficient overlap (Gelman & Hill, 2007) and lacked comparable counterfactuals (these students had propensity scores greater than .90). The analytic sample size for the main analysis was 574. Figure 1a displays the frequency distributions of propensity scores for treatment and control groups prior to conducting the one-to-one match. Figure 1b displays the frequency distributions of the propensity scores after conducting the one-to-one match.

We ran many models with interactions, transformed covariates, and the exclusion of collinear variables to achieve balance and obtain the appropriate specification for the propensity score model (Gelman & Hill, 2007). For binary variables, the criteria for strict balance was obtained if the difference in means was less than .05; for continuous variables, balance was achieved if there was less than a .1 standardized difference in means and a ratio of standard deviations between .85 and 1.15. There were substantial differences between the retained and promoted students on most covariates prior to matching. For example, the pool of promoted students had a higher proportion of females, whites, and students with married parents; these students also had higher reading scores, math scores, and teacher-reported academic skills (balance tables are available upon request). For the main analysis, we achieved balance on the majority (39 of 43) of the covariates and the investigators were satisfied with the final model specification. Only one of the covariates that failed to achieve balance was a higher priority covariate—the number of students in the school who were retained in the first grade. While this had a sound ratio of standard deviations, the standardized difference in means (.11) was slightly above the cutoff. The remaining three unbalanced covariates (binary variable for region,[6] continuous variable for percent of students of "other" race, and continuous variable for school average daily attendance), were determined a priori to be lower priority variables and close to the proposed cutoffs. To account for these unbalanced covariates, they were included in the regression-adjusted matched estimate (along with all high-priority variables).

[6]The authors considered region as one covariate, though it is organized using four dummy codes (Northeast, Midwest, South, and West). Balance for each dummy code was assessed separately, and Midwest and South were unbalanced.

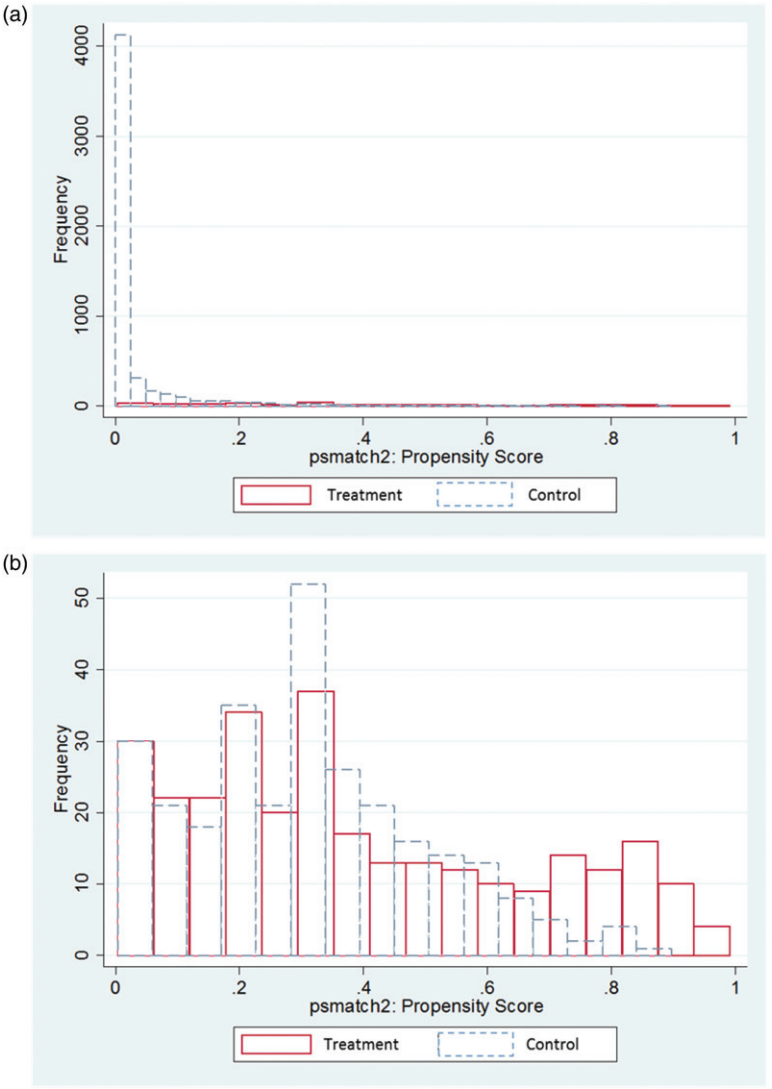

Figure 1. The overlap between the frequency distributions of the treatment and control groups before (a) and after (b) matching for the nearest neighbor propensity score approach.

Descriptive Results

Means and standard deviations for middle-grade (i.e., seventh- or eighth-grade) outcomes for treatment and control students identified using the nearest neighbor propensity score approach are displayed in Table 2. Retained students had an average reading IRT score of .89 ($SD = 0.31$), and promoted students had a mean score of .97 ($SD = 0.32$). Retained students had a mean math IRT score of .97 ($SD = 0.38$) and promoted students had a mean score of 1.00 ($SD = 0.40$). Means and standard deviations for the other academic and psychosocial outcomes were similar across the two groups. Reading and math test scores are positively

Table 2. Spring 2007 descriptive outcomes for the main analysis ($n = 574$).

	Retained ($n = 287$)		Promoted (n = 287)	
	Mean	*SD*	Mean	*SD*
Academic				
Reading IRT test score	0.89	0.31	0.97	0.32
Math IRT test score	0.97	0.38	1.00	0.40
Student self-reported reading competence[a]	2.34	0.73	2.34	0.73
Student self-reported math competence[a]	2.44	0.92	2.48	0.88
Teacher-report of reading competence: oral[b]	2.56	0.83	2.54	0.84
Teacher-report of reading competence: written[b]	2.13	0.84	2.15	0.82
Psychosocial				
Social self-concept[b]	3.64	0.97	3.45	0.98
Student self-reported internalizing behaviors[a]	2.17	0.61	2.17	0.61
Locus of control[c]	−0.35	0.67	−0.30	0.64
Self-esteem[c]	−0.23	0.69	−0.30	0.72

[a]Item response ranges from 1 to 4.
[b]Item response ranges from 1 to 5.
[c]Variable is a composite scale score with mean of zero and standard deviation of one.

correlated ($r = .67$, $p < .001$); correlations for all outcome variables are displayed in Table 3.

Estimated Treatment Effects

In the main analysis, we used the one-to-one nearest neighbor approach to identify a unique comparison student for each treatment student and then estimated the effect of retention on the set of outcomes. We found that grade retention had a small but statistically significant negative effect on middle school reading achievement ($b = -.06$, $p = .006$, $q = .055$; Table 4). The effect size measured in Cohen's d for reading achievement is −.19 (95% CI [−.35, −.02]). The negative direction of this standardized mean difference statistic (Durlak, 2009) confirms that the treatment of being retained in first or second grade had a negative impact on middle school reading achievement. In fact, being retained led the student to score .06 points lower on the standardized IRT reading test than if the child had not been retained. There were no detectable effects on math achievement or other academic outcomes (i.e., student self-reported reading and math competence, teacher-reported oral and written competence).

For psychosocial outcomes, there was a statistically significant positive coefficient for social self-concept ($b = .20$, $p = .014$, $q = .068$) indicating that retention led students to have slightly higher self-perceptions of their social acceptance and peer connectedness than if they were not retained. Cohen's d for social self-concept is .20 (95% CI [.04, .37]). No detectable effects were found for the remaining psychosocial outcomes (internalizing behaviors, locus of control, and self-esteem).

Sensitivity Analysis

To address concerns regarding omitted variable bias (e.g., second-grade test scores for the subsample of students, retained in second grade), we conducted a sensitivity analysis on the main analytic sample. According to the Wilcoxon signed rank test (Rosenbaum bounds p-critical = .05; Rosenbaum, 2002b), a confounding covariate increasing the

Table 3. Correlations among Spring 2007 outcomes for the main analysis ($n = 574$).

		1.	2.	3.	4.	5.	6.	7.	8.	9.
Academic	1. Reading IRT									
	2. Math IRT	0.67***								
	3. SR reading competence	0.12**	0.00							
	4. SR math competence	0.02	0.18***	0.18***						
	5. TR reading competence: oral	0.34***	0.36***	0.10*	0.01					
	6. TR reading competence: written	0.45***	0.45***	0.15***	0.04	0.75***				
Psychosocial	7. SR self-concept	0.01	−0.01	0.18***	0.14**	0.02	0.08*			
	8. SR internalizing behaviors	−0.18***	−0.24***	0.21***	0.00	−0.12**	−0.13**	−0.09*		
	9. SR Locus of control	0.34***	0.33***	0.09*	0.09*	0.20***	0.25***	0.13**	−0.37***	
	10. SR Self-esteem	0.24***	0.29***	0.21***	0.25***	0.15***	0.23***	0.35***	−0.38***	0.50***

***$p < .001$. **$p < .01$. *$p < .05$.

Note. SR: Student-report; TR: Teacher-report.

Table 4. Middle grade effects after grade retention: Findings from two propensity score approaches.

	Main analysis: One-to-one without replacement ($n = 574$)			Robustness check: Caliper with replacement ($n = 436$)		
	b	*SE*	CI	*b*	*SE*	CI
Academic outcomes						
Reading IRT test score	−0.06**	0.02	[−.10, −.02]	−0.05*	0.02	[−.10, −.01]
Math IRT test score	−0.01	0.02	[−.06, .03]	−0.03	0.03	[−.08, .03]
Student self-reported reading competence	−0.01	0.06	[−.13, .11]	0.05	0.08	[−.10, .21]
Student self-reported math competence	−0.04	0.07	[−.19, .11]	−0.06	0.09	[−.25, .12]
Teacher-report of reading competence: oral	0.06	0.07	[−.07, .19]	0.01	0.08	[−.15, .16]
Teacher-report of reading competence: written	0.01	0.06	[−.11, .13]	−0.08	0.08	[−.23, .07]
Psychosocial outcomes						
Social self-concept	0.20*	0.08	[.04, .35]	0.15	0.11	[−.06, .36]
Student self-report internalizing behaviors	−0.01	0.05	[−.11, .08]	−0.02	0.06	[−.14, .11]
Locus of control	−0.01	0.05	[−.11, .09]	−0.10	0.06	[−.21, .02]
Self-esteem	0.06	0.06	[−.05, .17]	0.05	0.08	[−.10, .19]

**$p < .01$. *$p < .05$.

odds of retention by 1.25 that nearly perfectly predicts the outcome (DiPrete & Gangl, 2004) would need to be present to eliminate the negative effect of retention on reading achievement.

Robustness Check: Caliper Matching With Replacement

To determine whether the outcomes from the main analysis were robust, we employed a different propensity score approach—caliper matching with replacement. The within-stratum treatment effect for retained individuals using the caliper matching approach was calculated using the following formula:

$$Y_i^T - (\Sigma Y_{ij}^C)/n_i \quad (3)$$

In Equation 3, the treatment unit stratum is represented via *i*, the matched control(s) within the stratum is *j*, and n_i represents the number of control students who are in that stratum. The individual treatment effects by stratum are summed and averaged to obtain the ATT. Following the recommendations of Rosenbaum and Rubin (1983), we set the caliper radius width to be .25 σ_p or one quarter of the standard deviation for the estimated propensity scores. In this study, matches did not differ by more than .03 in their propensity scores and multiple controls could be used for each treatment unit as long as they were within the specified distance. When estimating the treatment effect using this approach, propensity scores for controls were assigned frequency weights to account for the number of times they were used as matches for treatment students.

Retained students with propensity scores greater than .777 did not have matched control students within the given parameters and were dropped; therefore 43 treatment students were off common support. The analytic sample was smaller ($n = 436$) for this approach than the nearest neighbor main analysis because fewer treatment students had comparison students in the region of sufficient overlap (252 of 295 retained students), as the criteria for being a control student was more stringent (i.e., 184 control students

were within the .03 propensity score distance of treatment students). We achieved balance on 42 of 43 covariates; the only unbalanced variable was teacher-report of pre-retention student literacy (balance tables are available upon request). This variable had an acceptable standardized difference in means (−.098), but the ratio of the standard deviations (.80) was slightly below the desired threshold.

Results from the caliper matching approach (Table 4; obtained using covariate-adjusted linear regressions, Equation 2) are generally consistent with the main analysis (i.e., nearest neighbor) as evidenced by overlapping confidence intervals. In this robustness check, only reading achievement is significantly different from zero; the direction and magnitude of the estimate for reading achievement ($b = -.05$, $p = .028$) is similar to and corroborates our original finding from the main analysis. The effect size for this approach is also similar to the main analysis (Cohen's $d = -.18$). Using this method, the positive effect of social self-concept is no longer significant ($b = .15$, $p = .17$); therefore, this finding is not robust.

Discussion

Utilizing national data and propensity score methods, we generated causal estimates of the effects of first-grade or second-grade retention on academic and psychosocial outcomes six or seven years later. Retained students had statistically significant lower reading achievement in early adolescence compared to observably similar but promoted same-age peers. No other statistically significant effects were detected in both the main analysis and robustness check. Both propensity score models met strict balance criteria for the included and observed covariates, signaling that adequate comparisons were selected for the treatment students. Whereas most research focuses on either academic *or* psychosocial outcomes, and measures outcomes in elementary school, this study assesses the impact of early elementary grade retention on a range of student outcomes during the critical developmental period of early adolescence.

Although we do not find the average experience of early-grade retention to be detrimental across multiple developmental domains, neither do we detect positive effects of retention. The null findings do not prove that retained students would have had equivalent outcomes had they been promoted, but they also do not show retention to be successful. Unlike prior studies of early-grade retention that focused on specific regions and policies, we examined the impact of retention as a practice in a naturalistic context (e.g., retention services may vary by school, district, or state, and may or may not have included interventions such as summer school, intensive academic supports during the school year, etc.). The heterogeneity of implementation and receipt of retention as a treatment may contribute to the null findings. For instance, placing a student in the repeating year with the same teacher may yield negative effects, while requiring retained students to attend summer school and reassigning them to a high-quality teacher (as practiced in Florida; Winters & Greene, 2012) may yield positive estimates, producing a "net effect of zero." Thus in this study, we capture the average effect of retention across a wide range of remediation strategies that may be used across various schools, districts, and states, allowing us to more broadly understand retention's impact on students across the United States.

Interpreting the Negative Effect of Retention on Reading Achievement

On average, repeating first or second grade does not appear to give students the "gift of time" to acquire skills, but rather, triggers a more "negative cascade" and has a detrimental impact on subsequent reading achievement. This was supported by both the main analysis, which utilized a nearest neighbor approach with an adjustment for multiple comparisons, and also the caliper matching strategy (robustness check). Cohen's *d* was −.19 for reading test scores. Although convention suggests an effect size of .20 to be a "small" magnitude (Cohen, 1988), given that it comes from an evaluation of an educational intervention, this effect size is considered to have policy relevance (Hedges & Hedberg, 2007). Hong and Yu (2007) obtained comparable estimates of a quarter of a standard deviation lower for first-grade retainees on reading outcomes in 2004, when promoted students were in the fifth grade.

Although the reasons for retention in early elementary school are variable, low reading performance is a primary predictor of retention (Alexander et al., 2003) and can be a requirement for promotion (e.g., Özek, 2015). Low-performing readers are more likely to be held back in the same grade and be exposed to the same level of reading materials and type of pedagogical practices. Thus, they continue to struggle on reading assessments in the repeating year (Hong & Yu, 2007). Considering that low-performing readers may not receive additional, appropriate remediation, there are multiple reasons why retained students may have lagging reading skills into the middle grades. Unless individualized remedial efforts are implemented, the mere repetition of the grade will not be successful, thereby yielding persistent negative academic outcomes. Applying labeling theory (Becker, 1963), immediately after retention or later in elementary school, students may have internalized the belief that they are weak readers, which decelerates their longer term reading outcomes. Another explanation relates to expectancy theory, in which the lowering of elementary school teachers' expectations for the retained students' reading skills led to a self-fulfilling prophecy that contributed to slower rates of improvement over time (Weinstein, 2002). Although we did not detect significant effects on self- or teacher-reported reading competence *after* elementary school, these may have fluctuated over the years. Negative perceptions in late elementary school could have ended at the middle school transition, and may be a plausible mechanism for understanding negative reading achievement in the middle grades.

Contextualizing the Inconsistent and Null Effects of Retention

We did not detect significant effects for most outcomes, but this nonetheless contributes new and meaningful knowledge. Whereas we found a significant negative effect on reading achievement, none was detected for math achievement. If there were a bias due to the same-age comparison, one would expect there to be a negative effect of retention on math as well. The differential reading and math findings reveal that the consequences of retention may vary across subjects in the middle grades.

It is possible that this nonsignificant finding regarding math achievement may have to do with the uniqueness of math as a subject area. Generally, academic performance, self-esteem, and self-concept decrease after the transition to the middle grades; however, students may value math more and have greater interest in math than reading (Wigfield & Eccles, 1994). Thus, even if retained students are struggling in all subjects, there may

be greater effort put into math, yielding indistinguishable effects between those who were retained and promoted. That said, it is important to note that because retained students and their matched peers do not have statistically different math scores, it does not mean that retained students are succeeding in math. Both pools of students are low-performing as they were matched on academic performance prior to the treatment. They consistently remain low-performing when compared to the mean test scores of the full sample of promoted students in the spring of 2007 (Walston et al., 2008).

There was a small positive effect of early-grade retention on social self-concept in early adolescence in the main analysis, but this finding was not upheld in the robustness check. Social comparison theory (Festinger, 1954) may help explain this tenuous positive effect. Retained students completed these self-report surveys with their same-grade peers as their referent group. Given that retained students are older in age and potentially more socially advanced than their peers, this experience may lead to a "big fish little pond" effect as retained students feel more confident in their social skills than their promoted counterparts (Marsh & Craven, 2002). However, the effect was not corroborated in the caliper-matching approach and thus is not consistent. Although this may be due to issues of power with the smaller sample size that resulted from more stringent match criteria, it also suggests that this finding is less robust than it would need to be in order to interpret it with confidence. A handful of prior research studies find null or positive effects of retention on similar psychosocial outcomes in elementary school (e.g., Gleason et al., 2007; Wu et al., 2010). Results from the current study suggest that the post-retention boost in psychosocial outcomes are short-term at best as they are likely to go undetected in the long-term.

The null findings for teacher-reported oral and written competencies, student-reported academic outcomes, and the other psychosocial outcomes are important to consider. Although we might expect the findings for teacher-reported oral and written competencies to be similar to the findings for reading test scores (i.e., negative), teachers used a same-grade comparison group of the new, younger classmates in their reports of academic competence. The study's same-age comparison group (i.e., different grade level) of promoted students may obscure potential negative effects of retention on teacher-reported academic competence. In addition, prior work has found that first-grade retention increased behavioral engagement, as reported by teachers, up to three years post-retention (Wu et al., 2010). Although behavioral engagement is not studied here, classroom behaviors are strongly related to academic competence (Malecki & Elliot, 2002; Wentzel, 1991). These earlier positive effects are not corroborated, then, when examining a longer post-retention interval and a national sample of students. Finally, the current study's null findings for the psychosocial outcomes align with findings by Im et al. (2013), which report no differences in social-emotional trajectories of retained versus promoted students into eighth grade. These results counters isolated prior work that suggests the possibility of positive, short-term (i.e., elementary school), intrapersonal outcomes after retention (e.g., Gleason et al., 2007). Yet, the absence of negative effects does not provide evidence that retention is successful.

Limitations

There are several limitations worth noting. First, while we included a large number of covariates, it is impossible to satisfy the ignorability assumption. This is relevant for the

student self-reported psychosocial outcomes, as prior indicators were teacher- not student-reported. In the sensitivity analysis, the Wilcoxon signed rank test signals that gamma ($\Gamma = 1.25$) is not so large, making it not implausible that such an omitted variable exists. However, in this sample, only race and teacher-report of externalizing behavior significantly predicted retention with odds ratios greater than 1.25. Thus, it is unlikely that a confound related to the treatment that also nearly perfectly predicts the outcome is present given how predictive race and externalizing behaviors are in prior work. Alternatively, one could argue that the propensity score models were overfitted, leading to a loss of efficiency and increased variance (Chen et al., 2016). Future work examining the balance among reliability, efficiency, and precision is warranted, but for this sample, with the adjustment for multiple inference (FDR) and consistency of findings across two distinct propensity score matching strategies, we have reasonable confidence in these results.

Additionally, although balance was sound, it was not perfect. The means of the treated and untreated group prior to matching showed large differences between groups (i.e., retained students are more academically and psychosocially at-risk prior to retention). Although we attempted to address this bias, we were unable to do so for four confounding variables in the main analysis and included them as covariates when estimating the treatment effect.

Some limitations are related to the ECLS-K data set and construction of the treatment group. Though all students in this study had the same number of *years* in school, retained students had one less *grade* in school. Due to the ECLS-K's structure as a longitudinal study of a single cohort without data collection in adjacent years after the first grade, we are unable to examine both same-age and same-grade comparisons, which has only been done in isolated studies (e.g., Im et al., 2013). Further, we could not disentangle who was retained in first versus second grade due to the data-collection design of the ECLS-K. Thus, for the subset of second-grade retainees, data collected during second grade (including test scores) are omitted pretreatment confounders because they occurred prior to the retention decision. Finally, due to sample size limitations, we were unable to conduct subgroup analyses to illuminate potential heterogeneous effects of retention across groups.

Conclusion

Using a national data set and propensity score methods, the current study suggests that first- or second-grade retention lowers students' reading test scores six or seven years later. No other consistent and robust effects were found on psychosocial or academic outcomes. Yet, because these findings do not reveal retention as detrimental to a range of competencies, neither does it indicate that retention is effective, on average, for academically struggling students. The absence of compelling evidence to support this time- and resource-intensive intervention (Allen et al., 2009) indicates the need to systematically explore other educational alternatives, such as individualized tutoring, after-school and summer programs, and "social promotion plus" for struggling students (Jimerson, Pletcher & Kerr, 2005). In comparison to retention, these remediation strategies may yield better academic and psychosocial outcomes while also being more efficient. Ultimately, we hope this study motivates researchers, practitioners, and policymakers to identify ways

to better serve underperforming students in the early elementary years, so they may have stronger academic and psychosocial adaptation into and through the middle grades.

Acknowledgments

The authors wish to thank Erin Godfrey, Michael Kieffer, and Kate Schwartz for providing invaluable feedback during the design, analysis, and editing process.

Disclosure Statement

No potential conflict of interest was reported by the authors.

Funding

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B140037 to New York University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

ORCID

Sophia H. J. Hwang http://orcid.org/0000-0002-2341-6303
Elise Cappella http://orcid.org/0000-0002-3768-6443

ARTICLE HISTORY

Received 5 August 2016
Revised 23 June 2018
Accepted 25 June 2018

References

Adolph, K. E., & Robinson, S. R. (2008). In defense of change processes. *Child Development*, *79*(6), 1648–1653. doi:10.1111/j.1467-8624.2008.01215.x

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

Allen, C. S., Chen, Q., Willson, V. L., & Hughes, J. N. (2009). Quality of research design moderates effects of grade retention on achievement: A meta-analytic, multilevel analysis. *Educational Evaluation and Policy Analysis*, *31*(4), 480–499. doi:10.3102/0162373709352239

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), 1–20. doi:10.1300/J370v21n01_01

Anderson, L. W., Jacobs, J., Schramm, S., & Splittgerber, F. (2000). School transitions: Beginning of the end or a new beginning? *International Journal of Educational Research*, *33*(4), 325–339. doi:10.1016/S0883-0355(00)00020-3

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, *103*(484), 1481–1495. doi:10.1198/016214508000000841

Bali, V. A., Anagnostopoulos, D., & Roberts, R. (2005). Toward a political explanation of grade retention. *Educational Evaluation and Policy Analysis*, *27*(2), 133–155. doi:10.3102/01623737027002133

Balfanz, R. (2009). *Putting middle grades students on the graduation path: A policy and practice brief*. Westerville, OH: National Middle School Association.

Balfanz, R., Herzog, L., & Mac Iver, D. J. (2007). Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. *Educational Psychologist*, *42*(4), 223–235. doi:10.1080/00461520701621079

Bandura, K. E.. (2006). Toward a psychology of human agency. *Perspectives on Psychological Science*, *1*(2), 164–180. doi:10.1111/j.1745-6916.2006.00011.x

Becker, H. S. (1963). *Outsiders*. New York, NY: Free Press.

Becker, B. E., & Luthar, S. S. (2002). Social-emotional factors affecting achievement outcomes among disadvantaged students: Closing the achievement gap. *Educational Psychologist*, *37*(4), 197–214. doi:10.1207/S15326985EP3704_1

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300.

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, *93*(3), 491–507. doi:10.1093/biomet/93.3.491

Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review*, *15*(1), 1–40. doi:10.1023/A:1021302408382

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31–72. doi:10.1111/j.1467-6419.2007.00527.x

Cannon, J. S., & Lipscomb, S. (2011). *Early grade retention and student success: Evidence from Los Angeles*. San Francisco, CA: Public Policy Institute of California.

Cappella, E., Schwartz, K., Hill, J., Kim, H. Y., & Seidman, E. (2017). A national sample of 8th grade students: The impact of middle grade schools on academic and psychosocial competence. *The Journal of Early Adolescence*, *37*, 1–34. doi:10.1177/0272431617735653.

Chen, Q., Hughes, J. N., & Kwok, O. M. (2014). Differential growth trajectories for achievement among children retained in first grade: A growth mixture model. *The Elementary School Journal*, *114*(3), 327–353. doi:10.1086/674054

Chen, Q., Nian, H., Zhu, Y., Talbot, H. K., Griffin, M. R., & Harrell, F. E. (2016). Too many covariates and too few cases? A comparative study. *Statistics in Medicine*, *35*(25), 4546–4558. doi:10.1002/sim.7021

Clinton, W. J. (1998). *State of the union address (January 1, 1998)*. Washington, DC: U.S. Government Printing Office.

Cohen, J.. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: L. Erlbaum Associates.

DiPrete, T. A., & Gangl, M. (2004). Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Methodology*, *34*(1), 271–310. doi:10.1111/j.0081-1750.2004.00154.x

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology*, *34*(9), 917–928. doi:10.1093/jpepsy/jsp004

Eccles, J. S., & Midgley, C. (1989). Stage/environment fit: Developmentally appropriate classrooms for early adolescents. In R.E. Ames & C. Ames (Eds.). *Research on motivation in education* (Vol. *3*, pp. 139–186). San Diego, CA: Academic Press.

Eccles, J. S., Wigfield, A., Midgley, C., Reuman, D., Iver, D. M., & Feldlaufer, H. (1993). Negative effects of traditional middle schools on students' motivation. *The Elementary School Journal*, *93*(5), 553–574. doi:10.1086/461740

Eide, E. R., & Goldhaber, D. D. (2005). Grade retention: What are the costs and benefits? *Journal of Education Finance*, *31*(2), 195–214.
Elder, G. H. Jr. (1994). Time, human agency, and social change: Perspectives on the life course. *Social Psychology Quarterly*, *57*(1), 4–15.
Farrington, C. A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., & Beechum, N. O. (2012). *Teaching adolescents to become learners. The role of noncognitive factors in shaping school performance: A critical literature review*. Chicago, IL: University of Chicago Consortium on Chicago School Research.
Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*(2), 117–140. doi:10.1177/001872675400700202
Gelman, A., & Hill, J. (2007). *Data Analysis using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press.
Gleason, K. A., Kwok, O. M., & 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*(4), 327–340. doi:10.1086/516667
Gottlieb, K. G. (1992). *Individual development and evolution: The genesis of novel behavior*. New York: Oxford University Press.
Greene, J. P., & Winters, M. A. (2007). Revisiting grade retention: An evaluation of Florida's test-based promotion policy. *Education Finance and Policy*, *2*(4), 319–340. doi:10.1162/edfp.2007.2.4.319
Guo, S., & Fraser, M. W. (2010). *Propensity score analysis: Statistical methods and analysis*. Thousand Oaks, CA: Sage Publications.
Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response Theory*. Newbury Park, CA: Sage.
Heckman, J. J. (2000). Policies to foster human capital. *Research in Economics*, *54*(1), 3–56. doi:10.1006/reec.1999.0225
Heckman, J. J., LaLonde, R. J., & Smith, J. A. (1999). The economics and econometrics of active labor market programs. In A. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (Vol. 3, pp. 1865–2097). New York, NY: Elsevier Science.
Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87. doi:10.3102/0162373707299706
Hill, J. (2008). Discussion of research using propensity-score matching: Comments on "A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003" by Peter Austin, Statistics in Medicine. *Statistics in Medicine*, *27*(12), 2055–2061. doi:10.1002/sim.3245
Hill, J., & Reiter, J. P. (2006). Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine*, *25*(13), 2230–2256. doi:10.1002/sim.2277
Hill, J., Weiss, C., & Zhai, F. (2011). Challenges with propensity score strategies in a high-dimensional setting and a potential alternative. *Multivariate Behavioral Research*, *46*(3), 477–513. doi:10.1080/00273171.2011.570161
Hoffer, T. B. (1992). Middle school ability grouping and student achievement in science and mathematics. *Educational Evaluation and Policy Analysis*, *14*(3), 205–227. doi:10.3102/01623737014003205
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). New York, NY: 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*(3), 205–224. doi:10.3102/01623737027003205
Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy. *A Case Study of Causal Inference for Multi-Level Observational Data. Journal of the American Statistical Association*, *101*(475), 901–910.

Hong, G., & Yu, B. (2007). Early-grade retention and children's reading and math learning in elementary years. *Educational Evaluation and Policy Analysis*, *29*(4), 239–261. doi:10.3102/0162373707309073

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*(2), 407–421. doi:10.1037/0012-1649.44.2.407

Horton, N. J., & Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, *61*(1), 79–90. doi:10.1198/000313007X172556

Huang, F. L. (2014). Further understanding factors associated with grade retention: Birthday effects and socioemotional skills. *Journal of Applied Developmental Psychology*, *35*(2), 79–93. doi:10.1016/j.appdev.2013.12.004

Hughes, J. N., Cao, Q., West, S. G., Smith, P. A., & Cerda, C. (2017). Effect of retention in elementary grades on dropping out of school early. *Journal of School Psychology*, *65*, 11–27. doi:10.1016/j.jsp.2017.06.003

Hursh, D. (2007). Assessing No Child Left Behind and the rise of neoliberal education policies. *American Educational Research Journal*, *44*(3), 493–518. doi:10.3102/0002831207306764

Im, M. H., Hughes, J. N., Kwok, O. M., Puckett, S., & Cerda, C. A. (2013). Effect of retention in elementary grades on transition to middle school. *Journal of School Psychology*, *51*(3), 349–365. doi:10.1016/j.jsp.2013.01.004

Jackson, G. B. (1975). The research evidence on the effects of grade retention. *Review of Educational Research*, *45*(4), 613–635. doi:10.3102/00346543045004613

Jacob, B. A., & Lefgren, L. (2009). The effect of grade retention on high school completion. *American Economic Journal: Applied Economics*, *1*(3), 33–58. doi:10.1257/app.1.3.33

Jacob, B. A., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics*, *86*(1), 226–244. doi:10.1162/003465304323023778

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

Jimerson, S. R., & Ferguson, P. (2007). A longitudinal study of grade retention: Academic and behavioral outcomes of retained students through adolescence. *School Psychology Quarterly*, *22*(3), 314–339. doi:10.1037/1045-3830.22.3.314

Jimerson, S. R., Ferguson, P., Whipple, A. D., Anderson, G. E., & Dalton, M. J. (2002). Exploring the association between grade retention and dropout: A longitudinal study examining socio-emotional, behavioral, and achievement characteristics of retained students. *The California School Psychologist*, *7*(1), 51–62. doi:10.1007/BF03340889

Jimerson, S. R., Pletcher, S. M. W., & Kerr, M. (2005). Alternatives to grade retention. *Principal Leadership*, *5*(6), 11–15.

Karweit, N. L. (1999). *Grade retention: Prevalence, timing, and effects (CRESPAR Report No. 33)*. Baltimore, MD: Johns Hopkins University, Center for Social Organization of Schools.

Kim, H. Y., Schwartz, K., Cappella, E., & Seidman, E. (2014). Navigating middle grades: Role of social contexts in middle grade school climate. *American Journal of Community Psychology*, *54*(1–2), 28–45. doi:10.1007/s10464-014-9659-x

Lerner, R. M.. (2006). Developmental science, developmental systems, and contemporary theories of human development. In W. Damon & R. M. Lerner (Eds.), *The handbook of child psychology. Vol. 1: Theoretical models of human development* (6th ed., pp. 1–17). Hoboken, NJ: Wiley.

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

Liu, W., Kuramoto, S. J., & Stuart, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention Science*, *14*(6), 570–580. doi:10.1007/s11121-012-0339-5

Lorence, J. (2006). Retention and Academic Achievement Research Revisited from a United States Perspective. *International Education Journal*, *7*(5), 731–777.

Lorence, J. (2009). Grade retention redux: A dissenting perspective. In L. J. Saha & A. G. Dworkin (Eds.), *International handbook of research on teachers and teaching* (Vol. *2*, pp. 1153–1175). New York, NY: Springer.
Lorence, J. (2014). Third-grade retention and reading achievement in Texas: A nine year panel study. *Social Science Research*, *48*, 1–19. doi:10.1016/j.ssresearch.2014.05.001
Malecki, C. K., & Elliot, S. N. (2002). Children's social behaviors as predictors of academic achievement: A longitudinal analysis. *School Psychology Quarterly*, *17*(1), 1–23. doi:10.1521/scpq.17.1.1.19902
Mariano, L. T., & Martorell, P. (2013). The academic effects of summer instruction and retention in New York City. *Educational Evaluation and Policy Analysis*, *35*(1), 96–261. doi:10.3102/0162373712454327
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, J., Gershwin, D., Kirby, S., & Xia, N. (2009). *Retaining students in grade: Lessons learned regarding policy design and implementation*. Santa Monica, CA: RAND Corporation. Retrieved from http://www.rand.org/pubs/technical_reports/TR677.html
Masten, A. S., & Cicchetti, D. (2010). Developmental cascades. *Development and Psychopathology*, *22*(03), 491–495. doi:10.1017/S0954579410000222
Masten, A. S., & Coatsworth, J. D. (1998). The development of competence in favorable and unfavorable environments: Lessons from research on successful children. *American Psychologist*, *53*(2), 205–220. doi:10.1037/0003-066X.53.2.205
McCombs-Thomas, A., Armistead, L., Kempton, T., Lynch, S., Forehand, R., Nousianen, S., . . . Tannenbaum, L. (1992). Early retention: Are there long term beneficial effects? *Psychology in the Schools*, *29*(4), 342–347.
Moser, S. E., West, S. G., & Hughes, J. N. (2012). Trajectories of math and reading achievement in low-achieving children in elementary school: Effects of early and later retention in grade. *Journal of Educational Psychology*, *104*(3), 580–603.
Mulkey, L. M., Catsambis, S., Steelman, L. C., & Crain, R. L. (2005). The long-term effects of ability grouping in mathematics: A national investigation. *Social Psychology of Education*, *8*(2), 137–177.
Najarian, M., Pollack, J. M., & Sorongon, A. G. (2009). *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for the Eighth Grade (NCES 2009–002)*. National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.
National Association of School Psychologists (NASP). (2011). Grade retention and social promotion (Position Statement). Bethesda, MD. Retrieved from http://www.nasponline.org/about_nasp/positionpapers/GradeRetentionandSocialPromotion.pdf
No Child Left Behind (NCLB). (2002). Act of 2001, P. L. 107-110, § 115. *Stat, 1425*, 107–110.
Norwalk, K. E., Hamm, J. V., Farmer, T. W., & Barnes, K. L. (2016). Improving the school context of early adolescence through teacher attunement to victimization effects on school belonging. *The Journal of Early Adolescence*, *36*(7), 989–1009.
Özek, U. (2015). Hold back to move forward? Early grade retention and student misbehavior. *Education Finance and Policy*, *22*(3), 350–377. doi:10.1162/1162/EDFP_a_00166
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*(2), 297–315.
Prinstein, M. J., & Dodge, K. A. (2008). *Peer influence processes among youth*. New York, NY: Guilford.
Reardon, S. F., Cheadle, J. E., & Robinson, J. P. (2009). The effect of Catholic schooling on math and reading development in kindergarten through fifth grade. *Journal of Research on Educational Effectiveness*, *2*(1), 45–87. doi:10.1080/19345740802539267
Reschly, A. L., & Christenson, S. L. (2013). Grade retention: Historical perspectives and new research. *Journal of School Psychology*, *51*(3), 319–322.

Roderick, M., & Nagaoka, J. (2005). Retention under Chicago's high-stakes testing program: Helpful, harmful, or harmless? *Educational Evaluation and Policy Analysis*, *27*(4), 309–340. doi:10.3102/01623737027004309

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*(1), 123–158. doi:10.1207/s15327795jra0801_6

Rosenbaum, P. R. (2002a). Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, *17*(3), 286–327. doi:10.1214/ss/1042727942

Rosenbaum, P. R. (2002b). *Observational studies* (2nd ed.). New York, NY: Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. doi:10.1093/biomet/70.1.41

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, *6*(1), 34–58. doi:10.1214/aos/1176344064

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, *95*(450), 573–585. doi:10.1080/01621459.2000.10474233

Shepard, L. A., & Smith, M. L. (1988). Escalating academic demand in kindergarten: Counterproductive policies. *The Elementary School Journal*, *89*(2), 135–145. doi:10.1086/461568

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research*, *60*(3), 471–499. doi:10.3102/00346543060003471

Smith, M. L., & Shepard, L. A. (1988). Kindergarten readiness and retention: A qualitative study of teachers' beliefs and practices. *American Educational Research Journal*, *25*(3), 307–333. doi:10.3102/00028312025003307

Steinberg, L. (2008). *Adolescence* (8th ed.). Boston, MA: McGraw Hill.

Stuart, E. A., Marcus, S. M., Horvitz-Lennon, M. V., Gibbons, R. D., Normand, S.-L. T., & Brown, C. H. (2009). Using non-experimental data to estimate treatment effects. *Psychiatric Annals*, *39*(7), 414–451. doi:10.3928/00485713-20090625-07

Tomchin, E. M., & Impara, J. C. (1992). Unraveling teachers' beliefs about grade retention. *American Educational Research Journal*, *29*(1), 199–223. doi:10.3102/00028312029001199

Tourangeau, K., Nord, C., Lê, T., Pollack, J. M., & Atkins-Burnett, S. (2006). *Early Childhood Longitudinal Study–Kindergarten Class of 1998–99 (ECLS-K): Combined user's manual for the ECLS-K fifth grade data files and electronic codebooks (NCES Report No. 2006-032)*. Washington, DC: U.S. Department of Education.

Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., & Najarian, M. (2009). *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Combined User's Manual for the ECLS-K Eighth-Grade and K–8 Full Sample Data Files and Electronic Codebooks (NCES 2009–004)*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://nces.ed.gov/ecls/data/ECLSK_K8_Manual_part1.pdf

Walston, J., Rathbun, A., & Germino Hausken, E. (2008). *Eighth Grade: First Findings from the Final Round of the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (NCES 2008-088)*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington DC.

Warren, J. R., Hoffman, E., & Andrew, M. (2014). Patterns and trends in grade retention rates in the United States, 1995–2010. *Educational Researcher*, *43*(9), 433–443. doi:10.3102/0013189X14563599

Weinstein, R. S. (2002). *Reaching higher: The power of expectations in schooling*. Cambridge, MA: Harvard University Press.

Wentzel, K. R. (1991). Relations between social competence and academic achievement in early adolescence. *Child Development*, *62*(5), 1066–1078.

Wigfield, A., & Eccles, J. S. (1994). Children's competence beliefs, achievement values, and general self-esteem change across elementary and middle school. *The Journal of Early Adolescence*, *14*(2), 107–138. doi:10.1177/027243169401400203

Willson, V. L., & Hughes, J. N. (2009). Who is retained in first grade? A psychosocial perspective. *The Elementary School Journal*, *109*(3), 251–266. doi:10.1086/592306

Winters, M. A., & Greene, J. P. (2012). The medium-run effects of Florida's test-based promotion policy. *Education Finance and Policy*, *7*(3), 305–330.

Wu, W., West, S. G., & Hughes, J. N. (2008). 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*(4), 727–740. doi:10.1037/a0013098

Wu, W., West, S. G., & Hughes, J. N. (2010). Effect of grade retention in first grade on psychosocial outcomes and school relationships. *Journal of Educational Psychology*, *102*(1), 135–152. doi:10.1037/a0016664

U.S. Department of Education. (2009). *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K) Kindergarten through Eighth Grade Full Sample Public-Use Data and Documentation (DVD) (NCES 2009–005)*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Xia, N., & Kirby, S. (2009). Retaining students in grade: A literature review of the effects of retention on students' academic and nonacademic outcomes (Technical Report TR-678-NYCDOE). Retrieved from http://www.rand.org/pubs/technical_reports/2009/RAND_TR678.pdf

Appendix A

Table A1. Selected Covariates for Propensity Score Development (Organized by Priority)

Higher priority covariates	Lower priority covariates
Gender	Region[b]
Race[a]	Urbanicity[c]
Parent marital status	Father employment
Student has disability at start of kindergarten	Mother employment
SES	Family in poverty
Age of child in months (spring of first grade)	Number of people in household
First-grade reading test score	Mother education
First-grade math test score	Student has disability in spring of first grade
First-grade general knowledge test score	Teacher-report of interpersonal skills
Teacher-report externalizing problems	Teacher-report of self-control
Teacher-report approaches to learning	Teacher-report of internalizing problems
Teacher-report math competence	School average daily attendance
Teacher-report literacy competence	% of students in school who are racially "other"
Number of students enrolled in first grade	% of limited English proficient students in school
% in school at or above grade level in reading	% of limited English proficient students in first grade
% in school at or above grade level in math	% eligible in school for free lunch
% of Hispanic students in school	If student can be retained in any grade[d]
% of Black students in school	If student can be retained by parent request[d]
School has formal retention policy	If student can be retained because of below- grade level performance[d]
Number of retained first graders in school	If student can be retained in kindergarten[d]
	If student can be retained in any grade more than once[d]
	If student with disability can be retained[d]
	If student can be retained without parent approval[d]

[a]Organized using the following dummy-coded variables: White, Black, Hispanic, Asian, Native American/Pacific Islander, Multiracial.
[b]Organized using the following dummy-coded variables: Northeast, Midwest, South, West.
[c]Organized using the following dummy-coded variables: urban, suburban, rural.
[d]Specific school retention policies.