Developmental Psychology
2008, Vol. 44, No. 2, 407–421

Copyright 2008 by the American Psychological Association
0012-1649/08/$12.00 DOI: 10.1037/0012-1649.44.2.407

Effects of Kindergarten Retention on Children's Social-Emotional Development: An Application of Propensity Score Method to Multivariate, Multilevel Data

Guanglei Hong and Bing Yu
Ontario Institute for Studies in Education of the University of Toronto

This study examines the effects of kindergarten retention on children's social-emotional development in the early, middle, and late elementary years. Previous studies have generated mixed results partly due to some major methodological challenges, including selection bias, measurement error, and divergent perceptions of multiple respondents in different domains of child development. The authors address these challenges by using propensity score stratification to contend with selection bias and by embedding measurement models in hierarchical models to account for measurement error and to model dependence among observations. The authors' analyses of a series of multivariate models enable them to compare the retention effects across different respondents over different time points. In general, the results show no evidence suggesting that kindergarten retention does harm to children's social-emotional development. Rather, the findings suggest that, had the retained kindergartners been promoted to the first grade instead, they would possibly have developed a lower level of self-confidence and interest in reading and all school subjects 2 years later and would have displayed a higher level of internalizing problem behaviors at the end of the treatment year and 2 years later.

Keywords: causal inference, grade retention, measurement error, nonexperimental data, sensitivity analysis

Supplemental materials: http://dx.doi.org/10.1037/0012-1649.44.2.407.supp

With a growing emphasis on national standards and accountability, grade retention has become increasingly common in the United States. According to the National Association of School Psychologists, the kindergarten through 12th-grade retention rate increased by 40% over 20 years (Dawson, 1998). By 1995, the annual rate of retention rose to 13.3% (U.S. Bureau of the Census, 1995). In the meantime, the kindergarten retention rate was about 6% in 1993 and 5% in 1995 (Zill, Loomis, & West, 1997). The debate about the benefits and consequences of grade retention ensues in this context. Past research has often been inconclusive with regard to the effects of grade retention on child development (Holmes, 1989; Holmes & Matthews, 1984; Jimerson, 2001). Even more controversial is the practice of retaining young children in kindergarten (Shepard & Smith, 1989).

Retention in a higher grade is often resorted to as a remedy for students who are behind academically. In contrast, at the kindergarten level, many children are retained for behavioral rather than academic reasons. Teachers and parents perceive these children to be socially or emotionally immature in adapting to the school environment and therefore not ready for first-grade learning (Byrnes, 1989). Results from recent studies using a large-scale longitudinal data set suggested that retained kindergartners would have learned more in reading and mathematics at the end of the treatment year had they been promoted to the first grade instead (Hong & Raudenbush, 2005, 2006). Although the negative effects of retention on the retained students' reading and math achievement seemed to diminish over years, researchers found no evidence that kindergarten retention brought a general advantage to the retained students' cognitive learning 2 and 4 years after the treatment (Hong & Yu, 2007). A question remaining unanswered is whether kindergarten retention is beneficial to the retained students' social-emotional development over their elementary years.

Analyzing a nationally representative sample from the United States, we aim, in the present study, to investigate the relationships between kindergarten retention and children's social-emotional outcomes. We examine multiple domains, including children's self-perceived competence and interest in reading, math, and all school subjects; their self-reported competence and interest in peer

Guanglei Hong and Bing Yu, Department of Human Development and Applied Psychology, Ontario Institute for Studies in Education of the University of Toronto, Toronto, Ontario, Canada.

The names of the authors are in alphabetical order. This research has received support from a Social Sciences and Humanities Research Council of Canada Institutional Grant and a National Academy of Education/Spencer Postdoctoral Fellowship awarded to Guanglei Hong. Additional support came from a major research grant titled "Improving Research on Instruction: Models, Designs, and Analytic Methods" funded by the Spencer Foundation. Opinions reflect those of the authors and do not necessarily reflect those of the granting agencies.

We thank Jenny Jenkins, Christine Li-Grining, Stephen Raudenbush, and Sean Reardon for helpful suggestions.

Correspondence concerning this article should be addressed to Guanglei Hong, Ontario Institute for Studies in Education of the University of Toronto, Toronto, Ontario, Canada M5S 1V6. E-mail: ghong@oise.utoronto.ca

relationships; and internalizing problem behaviors as perceived by teachers, parents, and the children themselves. We ask the following question: If a child at risk of repeating kindergarten is actually retained, how would the child develop in the above social-emotional domains at the end of the treatment year, 2 years later, and 4 years later in comparison with the expected outcomes of being promoted to the first grade?

Previous Research

Retention, especially kindergarten retention, has long been a theoretically controversial issue. Proponents of early intervention argue that retention can be beneficial in the primary grades because it will prevent failures in academic learning and in social-emotional development from occurring or becoming severe (Shepard & Smith, 1989). From this perspective, rather than allowing social-emotional problems to impede children's later development, an additional year in kindergarten, "a gift of time," will enable children to reach the maturity level required for adapting to the school setting. Hence, retained kindergartners are expected to have an increased chance of success when they take on more advanced learning tasks and face more complex environmental challenges in the later grades. The early intervention theory therefore provides a justification for the practice of retaining kindergartners who are considered to be relatively behind their classmates socially and emotionally. The theory predicts a positive effect of kindergarten retention on the retained students' short-term and long-term success in all domains of development. From the opposing point of view, however, one may argue that retention deprives children of opportunities to engage in age-appropriate cognitive and social activities and therefore may suppress their academic interest and interrupt their development of self-regulation and interpersonal skills (Morrison, Griffith, & Alberts, 1997). As a result, retained kindergartners may lag further behind their same-age peers in the later years.

Among various domains of social-emotional outcomes, children's self-perceived competence in academic subjects and in peer relationships and their internalizing problem behaviors are considered to be particularly sensitive to the retention intervention. Different theories point in different directions in terms of the retention effects on child development in these domains. In general, a child develops self-identity through perceiving his or her relative standing among the proximate peers. However, there are contrasting theoretical arguments about how the change in peer composition as a result of retention may affect the retained student's self-concept.

Social comparison theory (Festinger, 1954) has supplied a rationale for grade retention. During the repetition year, retained kindergartners not only have the second exposure to the most basic content in academic subjects, but they also find themselves among a group of younger peers who have never received formal education. Having already obtained some academic skills from their first kindergarten year, the retained children will likely feel a higher level of competence in the academic subjects when compared with their new classmates. In addition, having accumulated a whole year of experience in socializing with peers and coping with conflicts in school settings, the retained children may appear to be more knowledgeable and competent in peer relationships in comparison with the first-time kindergartners (Plummer & Graziano, 1987). Following this line of thinking, spending a second year in kindergarten may improve the retained students' self-perceived competence in school subjects as well as in interpersonal relationships. The social comparison perspective is often endorsed by teachers and parents who are primary decision makers in the retention process. They expect that retention may become a turning point for the retained students as they gain academic and social advantages among a group of younger children (Byrnes, 1989; Tomchin & Impara, 1992).

In contrast, inferring from the labeling theory (Becker, 1963; Lemert, 1967), opponents predict a possible detrimental effect of retention on children's development in self-esteem and self-perceived competence. When retention is associated with certain labels carrying negative meanings such as "incompetence" or "deviance," retained students may interpret being retained in a grade as being rejected by their teachers and same-age peers and thus feel humiliated and discouraged (Pagani, Tremblay, Vitaro, Boulerice, & McDuff, 2001). Byrnes's (1989) study with children in Grades 1 to 6 found that children in grades as low as Grade 1 were capable of understanding the concept of retention and viewed it as a punishment. When being interviewed, about one quarter of the retained students denied the fact that they had been retained. Both retained children and promoted children described being retained as a "sad," "bad," and "upsetting" experience. Moreover, young children are particularly sensitive to age differences (Morrison & Perry, 1956). Retained students, often overaged among their new peers, tend to feel alienated and may choose to withdraw from social activities. Hence, children being labeled as "retainees" will likely develop internalizing problem behaviors such as feeling sad and lonely and having low self-esteem. In school subjects, repeating the same curriculum may bore the retained students and may bring back to them the frustrating experiences they had when they first encountered the same learning content. Moreover, the labeling theory predicts that being labeled as "slow learners" or "low achievers" will lead to low self-expectations and low self-efficacy in academic learning.

However, the labeling theory has received challenges from a counterargument claiming that kindergartners are perhaps too young to be capable of processing conflicting social evaluations. Studies have shown that the retention effects on children's social-emotional development may differentiate by grade level, with early retention bearing less of a negative impact than later retention on children's self-concepts (Finlayson, 1977; Morrison & Perry, 1956; Tomchin & Impara, 1992). As a result, retained kindergartners may not suffer greatly from the stigma typically associated with retention (Shepard, 1989).

Past research has generated mixed results regarding the effects of grade retention on social-emotional development, mostly showing insignificant effects (Pierson & Connell, 1992; Shepard & Smith, 1989). Some studies reported positive effects of grade retention on children's self-esteem and self-perceived competence. For example, Holmes's (1989) meta-analysis reported a small positive effect (effect size = .06) on self-concept on the basis of six studies. Using a sample of low-income, mostly Black children in the Chicago Longitudinal Study, Reynolds (1992) found that retention was positively related to children's self-perceived school competence, especially for early-retained children. However, this early intervention failed to produce sustainable positive effects in the long run. Analyzing the same sample several years later, McCoy and Reynolds (1999) found that the positive effect vanished by age 14.

On the contrary, Holmes and Matthews's (1984) meta-analysis concluded an effect size of −.02 of the retention effect on self-concept. Morrison and Perry (1956) investigated children's self-perceived peer relationships by comparing overage children with at-age and underage children in the same grade. They found that overage children, mostly retained in grade, perceived a significantly lower level of acceptance by peers. In the Minnesota Mother–Child Interaction Project, researchers compared a group of children who had been retained once from kindergarten through Grade 3 with a low-achieving promoted group and a control group randomly selected from all the promoted children. The retained children were ranked the lowest on self-esteem and peer acceptance and highest on behavior problems immediately after retention. The between-groups differences were sustained up to age 16 (Jimerson, Carlson, Rotert, Egeland, & Sroufe, 1997). Analyzing data from the Quebec Longitudinal Study of Kindergarten Children with adjustment for child sociodemographic characteristics, kindergarten inattentiveness, and a random effect for each child, Pagani et al. (2001) reported increasingly persistent anxious, inattentive, and disruptive behaviors after grade retention in primary school.

Given the contradictory evidence from the past research, the debate about whether kindergarten retention promotes or impedes children's social-emotional development remains unsettled. In order to further theoretical understanding and to inform practice, we focused our study on testing the contrasting hypotheses derived from the alternative theoretical perspectives. To be specific, the early intervention theory and the social comparison theory predict that (a) retained kindergartners would gain more self-efficacy and develop more interest in academic learning than they would have if promoted, and (b) retained kindergartners would show more competence in peer relationships and enjoy more popularity than they would have if promoted. The labeling theory, in contrast, hypothesizes that (c) retained kindergartners would display more internalizing problem behaviors than they would have if promoted.

We define the causal effect of retention versus promotion for a child attending a certain school as the difference between the child's potential outcome if retained and his or her potential outcome if promoted in a school of the same type. This causal effect is defined under a weak version of the stable-unit-treatment-value assumption that there is a single value of each potential outcome associated with each treatment for each child given the school setting in which the treatments are carried out (Hong & Raudenbush, 2006; Rubin, 1986). Our theoretical interest is in the population average retention effect. We may consider three populations as possible targets of inference: (a) all children, (b) all children at risk of repeating kindergarten, and (c) retained children. Because kindergarten retention is highly selective, many children in the first population have almost no risk of ever being retained in kindergarten under the current system. To these children, the theoretical questions that we have raised above have little relevance. The causal effects of retention versus promotion can nonetheless be defined for both Population b and Population c. When these two populations have different compositions and when the retention effect is not constant, the average retention effect for Population b will likely be different from that for Population c. To simplify, we focus our current study on evaluating the average retention effects for the population of children at risk of repeating kindergarten. We discuss in the final section how to investigate heterogeneous retention effects and how to estimate the retention effects for the population of retained children.

Methodological Issues

Empirical results from many previous retention studies have not been highly informative to theories and to policy making not only because the findings have been mixed. More important, some major methodological challenges have hindered attempts at evaluating the retention effects. These challenges include selection bias, measurement error, and divergent perceptions of multiple respondents in multiple domains of child development. Below we discuss each of these problems and introduce some promising analytical solutions. Specifically, we suggest using propensity score matching or stratification to adjust for a very large number of pretreatment predictors of retention. We explain how to embed measurement models in hierarchical models to account for measurement errors and to adjust for dependence among observations. In addition, we adopt multivariate, multilevel models to reveal the similarities or differences in the retention effects across different respondents and across different time points. We illustrate these methods in our application study in the next section.

Selection Bias

Because it is impractical to conduct a large-scale experiment assigning children at random to be retained, researchers must rely on nonexperimental data in addressing the causal question of whether the retained students actually experienced better cognitive and social-emotional development than they would have had they been promoted instead. The essential problem here is how to identify an appropriate comparison group for the retained group. Previous studies have compared the outcomes of the retained children either with their new classmates who were experiencing the grade for the first time (i.e., same-grade comparison) or with their same-age peers who had been promoted to the next grade (i.e., same-age comparison). Obviously, in same-grade comparisons, most retained students cannot be matched with their new classmates on age. Hence, the comparison group provides little counterfactual information about how the retained students would fare had they been promoted. Even in same-age comparisons, the retained group and the promoted group are still vastly different on average in many prior characteristics. Adjustment by means of a linear model or multivariate matching sharply constrains the number of background variables that can be controlled (Little, 1985; Stone, 1993). In particular, most promoted children have little or no risk of ever being retained. When the two groups are barely comparable, statistical adjustment for a limited number of background variables cannot be relied upon to remove bias. Moreover, predictions about what might happen to the retained children if promoted are based largely on linear "extrapolations" without the support of empirical data (Little, An, & Johanns, 2000; Lord, 1967; Shadish, Cook, & Campbell, 2002).

Propensity Score Adjustment

In this study, we use a propensity score to adjust for about 200 background variables that may potentially confound the retention effect estimation. Including all these covariates in a multiple regression or an analysis of covariance will greatly reduce the degrees of freedom left for estimating the retention effects. We solve this problem by using a unidimensional propensity score that summarizes all the information that predicts retention. The propensity score indicates a child's propensity of being retained and can be estimated as a function of all the observed pretreatment covariates. As proved by Rosenbaum and Rubin (1983), subsets of retained and promoted children who have the same propensity score should have the same joint distribution of all the observed pretreatment covariates. For example, suppose that we estimate the propensity score as a function of age, gender, and socioeconomic status. The retained group and the promoted group that have the same propensity score should have the same age, gender, and socioeconomic status compositions. Hence, statistical adjustment for the propensity score should be sufficient for removing the selection bias associated with these observed covariates. Rosenbaum (1987) showed that adjusting for estimated propensity scores corrects for both systematic and chance imbalances in the observed covariates in a finite sample and therefore often outperforms adjustment for true propensity scores.

Propensity score-based adjustment methods include the following: (a) Use nearest neighbor matching to identify control units that can be matched with the treated units on the estimated propensity score when there is a large reservoir of control units (for application examples, see Hill, Waldfogel, Brooks-Gunn, & Han, 2005; Rosenbaum, 1986). (b) Stratify the sample on the estimated propensity score, estimate within-stratum treatment effects, and generate an overall average treatment effect estimate by pooling the results over all the strata (for application examples, see Hong & Raudenbush, 2005, 2006; Hong & Yu, 2007; Rosenbaum & Rubin, 1984). According to Cochran (1968), stratifying a sample into five groups on the basis of a pretreatment covariate will remove about 90% of the selection bias associated with this covariate. (c) Adjust for the estimated propensity score as a covariate under model-based assumptions when such assumptions seem plausible. (d) Use the estimated propensity score to compute a sample weight proportionally inverse to one's conditional probability of receiving the treatment that one actually received. Examples include inverse-probability-of-treatment weighting (Robins, 2000; Robins, Hernan, & Siebert, 2003) and marginal mean weighting (Hong, 2007; Hong & Hong, 2007; Huang, Frangakis, Dominici, Diette, & Wu, 2005; Imbens, 2000). Propensity score adjustment cannot remove bias associated with unmeasured confounders that are independent of the observed covariates. Hence, the above-described methods will generate unbiased estimates of the treatment effects only if the treatment assignment is not associated with unmeasured covariates given the observed covariates. This is the so-called strong ignorability assumption in the causal inference literature (Rosenbaum, 1984; Rosenbaum & Rubin, 1983). For this reason, propensity score adjustment is especially effective when researchers have collected a comprehensive list of pretreatment covariates. The robustness of the estimation results in the presence of potential unmeasured confounders can be assessed through sensitivity analysis.

Dependence Among Observations

Because grade retention occurs in school settings, children attending the same school are influenced together by the school context. The probability that a child is retained in a certain school is partly dependent on who else is attending the same school. By the same token, the social-emotional outcomes of a child are usually not independent of the social-emotional outcomes of his or her schoolmates. In studies that involve repeated observations of a child, the assumption of independent observations, typically invoked in standard regressions or analyses of covariance, is again violated. Statistical analyses that ignore dependence among observations may misestimate the sampling error and may consequently produce misleading results in hypothesis testing. Multilevel modeling enables us to obtain valid analytic results through appropriate adjustment for the dependence among observations. In addition, by adopting a weak version of the stable-unit-treatment-value assumption that reflects the nested structure of the data, it becomes possible to define and estimate school-specific retention effects. Specifically, we estimate every child's propensity for being retained through analyzing a hierarchical logistic regression model with children nested within schools. We then estimate the retention effects on children's social-emotional outcomes in a hierarchical linear model, which has the capacity of assessing the variation in the retention effects across schools that may be associated with school composition or treatment implementation.

Measurement Error

Social-emotional outcomes tend to be elusive and prone to measurement error. Measurement error in an outcome will add noise to the residual variance, reduce the magnitude of the estimated effect size, and attenuate the correlations among different outcomes. Measurement error in a predictor is more consequential, because it will bias not only the coefficient estimate of the error-laden predictor but also the coefficient estimates of other predictors in the same model that are correlated with this predictor. Most previous studies of retention have developed survey questionnaires or adopted instruments with known psychometric properties. For example, Mantzicopoulos and Morrison (1992) reported alpha reliabilities ranging from .68 to .94 for the six subscales measured with the Revised Behavior Problem Checklist constructed by Quay and Peterson (1987). However, past researchers have not used such psychometric information in statistical analyses when comparing the social-emotional outcomes between the retained group and the comparison group. In our analysis, we use reliability information to account for measurement errors in both predictors and outcomes. Specifically, through adjusting for the estimated true score of each pretest in the outcome models, we manage to reduce bias and in the meantime improve precision in treatment effect estimation. In addition, having obtained the error variance computed from the reliability of each outcome measure, we are able to model the dependence among the true scores of multiple outcomes at the child level and the school level and obtain retention effect estimates with improved precision.

Multivariate Outcomes

In the current study, measures of children's social-emotional development encompass a number of domains, including competence and interest in reading, math, and all subjects; competence and interest in peer relationships; and internalizing problem behaviors. Past studies of grade retention have typically used univariate analyses. That is, separate analyses were conducted for different outcomes. This conventional analytic strategy not only contains inflated Type I errors due to multiple hypotheses testing in a single sample, but it also fails to portray the inherent links among the social-emotional outcomes in different domains or subdomains, among different respondents, and across different time points. In our analysis, we develop a series of multivariate, multilevel models (Cheong & Raudenbush, 2000; Raudenbush, Brennan, & Barnett, 1995; Raudenbush & Bryk, 2002; Raudenbush, Rowan, & Kang, 1991). For a given set of outcome measures in each model, we examine correlations among the true scores at the child level and the school level. With repeated observations of a child's social-emotional behaviors over multiple years, we can possibly detect whether the retention effects in a given domain are sustained or diminishing in the long run. In addition, children's internalizing problem behaviors were rated by teachers, parents, and children themselves. Teachers usually compare a focal child with other children within the school setting. Parents' perceptions of a child's behaviors within the home setting provide a different lens. Once reaching a certain age, children themselves become capable of observing and reflecting upon their own social-emotional competencies and problems. An examination of the degree of consistency across teachers, parents, and children will enable us to obtain a relatively comprehensive evaluation of a child's developmental status. In cases in which multiple *t* tests lead to rejections of null hypotheses, we conduct an omnibus chi-square test for all the retention effects in the same model. With abundant within-school data, a multivariate multilevel model may also supply information about the variation and covariation of school-specific retention effects across different outcomes.

Method

Sample

The current study evaluates the short-term and long-term effects of kindergarten retention on children's social-emotional development through analyzing the Early Childhood Longitudinal Study Kindergarten cohort (ECLS-K) data released by the U.S. National Center for Education Statistics (NCES). Data were collected from a nationally representative sample of about 21,000 children, their parents, teachers, and schools in a total of six waves in fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004. The sample is composed of about 55% White, 15% Black, 18% Hispanic, and 6% Asian children, and 5% of the children in the sample were Native American, Hawaiian, or of other races. Nearly 20% of the children came from families living below the poverty line.

Measures

Treatment. Information about whether a child was retained in kindergarten in Year 1 (i.e., the 1999–2000 school year) was available for only about 50% of the children in the ECLS-K full sample. Earlier research has found that the retention schools were systematically different from the nonretention schools (Hong & Raudenbush, 2005). Our analytic sample therefore consists of 10,726 first-time kindergartners attending 1,080 schools that allowed for kindergarten retention. Among these kindergartners, 471 were retained, and the rest were promoted to the first grade. In comparison with the full sample, our analytic sample has a slightly lower proportion of poor children, minority children, and children from non-English-speaking families. Because those children whose treatment information was missing also tended to miss information on many other variables, we were unable to compute a new sample weight for each unit in our analytic sample.

Posttreatment self-perceived competence and interest in school learning. In the springs of Year 3 (i.e., the 2001–2002 school year) and Year 5 (i.e., the 2003–2004 school year), every child was asked to rate on a scale of 1–4 his or her self-perceived competence, difficulties, interest, and enjoyment in reading (eight items), mathematics (eight items), and all school subjects (six items). We used the mean rating of each set of items in each year as the outcome measure. The reliabilities of these six outcomes range from .79 to .92.

Posttreatment self-perceived competence and interest in peer relationships. Also in the springs of Year 3 and Year 5, every child was asked to respond to a set of six items measuring, on a scale of 1–4, how easily he or she made friends, got along with other children, and had popularity among peers. The reliabilities are .79 in Year 3 and .82 in Year 5. The above measures were adapted from the Self-Description Questionnaire I (Marsh, 1990).

Posttreatment internalizing problem behaviors. Teachers and parents reported, on a scale of 1–4, the apparent presence of sadness, loneliness, and low self-esteem in a child. Every child received teacher ratings in the springs of Year 1, Year 3, and Year 5 and a parent rating in the spring of Year 1. The teacher Social Rating Scale and the parent Social Rating Scale, each consisting of four items for measuring internalizing problem behaviors, were adapted from the Social Skills Rating Scale: Elementary Scale A (Gresham & Elliott, 1990). The reliabilities of the teacher ratings are around .77; the parent rating shows a lower reliability ($\lambda =$ .63). In addition, every child responded to eight items, on a scale of 1–4, measuring his or her own feelings of sadness, loneliness, and low self-esteem in the springs of Year 3 and Year 5. The child measures, adapted from the Self-Description Questionnaire I, show a reliability of .81 in Year 3 and .79 in Year 5 (NCES, 2002; Pollack, Atkins-Burnett, Tourangeau, & West, 2005; Pollack, Najarian, Rock, Atkins-Burnett, & Hausken, 2005; see Table 1 for information on these scales). In most cases, large-scale surveys cannot afford extensive measurement of a single psychological trait such as internalizing problem behaviors. The limited number of items in teacher and parent ratings may have failed to fully capture a child's well-being and may have led to the relatively low reliabilities of these measures.

Reading and math pretest scores. At the end of Year 0 (i.e., spring 1999), every child was assessed on about 50–70 items in each subject area, including reading and mathematics. The NCES researchers used a three-parameter item response theory (IRT) model (Hambleton, Swaminathan, & Rogers, 1991) to estimate a child's latent ability in each of these subjects. The average reliabilities were .95 for the reading IRT scores and .94 for the math IRT scores. In our data, these two pretest scores are the most important predictors of child self-perceived competence and interest in academic learning 2 and 4 years after the treatment.

Pretreatment teacher- and parent-rated interpersonal relationships. At the end of Year 0, every child received teacher and parent ratings on interpersonal relationships. The teacher rating was based on five items with a reliability of .89. The parent rating

Table 1
Descriptions of Domains in Children's Social-Emotional Development

Measure, respondent, and year	Description	No. of items	Reliability[a]	*M*	SD
Perceived interest and competence in academic learning					
Child	Reading grade, the difficulty of reading work, and interest in and enjoyment of reading	8			
Year 3			.87	3.27	0.65
Year 5			.90	3.00	0.72
Child	Mathematics grade, the difficulty of mathematics work, and interest in and enjoyment of mathematics	8			
Year 3			.90	3.16	0.78
Year 5			.92	2.92	0.77
Child	Performance in all school subjects, and enjoyment of all school subjects	6			
Year 3			.79	2.92	0.64
Year 5			.83	2.73	0.63
Perceived interest and competence in peer relationships					
Child	How easy it is to make friends and get along with children, self-perception of popularity	6			
Year 3			.79	3.03	0.63
Year 5			.82	3.00	0.59
Internalizing problem behaviors					
Child	Feeling sad a lot of the time, feeling lonely, feeling ashamed of mistakes, and worrying about school and friendships	8			
Year 3			.81	2.14	0.72
Year 5			.79	2.00	0.61
Teacher	Apparent presence of anxiety, loneliness, low self-esteem, and sadness	4			
Year 1			.77	1.58	0.51
Year 3			.76	1.61	0.52
Year 5			.77	1.62	0.54
Parent	Apparent problems with being accepted and liked by others, sadness, loneliness, and low self-esteem	4			
Year 1			.63	1.53	0.39

[a] Split-half reliability for teacher and parent ratings; alpha coefficient for child self-rating.

was based on only three items with a reliability of .68. We found these two pretreatment ratings to be the most important predictors of child self-perceived competence and interest in peer relationships 2 and 4 years after the treatment.

Pretreatment teacher- and parent-rated internalizing problem behaviors. Also at the end of Year 0, every child received teacher and parent ratings on internalizing problem behaviors. The teacher and parent ratings contained four items each, with reliabilities of .78 and .61, respectively. These two ratings are the most important pretreatment predictors of posttreatment teacher ratings, parent ratings, and child self-ratings of internalizing problem behaviors.

Analytic Procedure

Our analysis of the causal effects of kindergarten retention involves five major steps. In Step 1, we estimate every child's propensity of being retained as a function of the observed pretreatment covariates. In Step 2, we stratify the sample of children on the basis of the logit of the estimated propensity score such that the retained children and the promoted children in the same propensity stratum show the same distributions in almost all the observed pretreatment covariates. The within-stratum mean difference in a social-emotional outcome between the retained children and the promoted children estimates the retention effect on these children. In order to remove residual selection bias, if there is any, associated with the strongest predictors for each set of outcomes, we make additional adjustments for these pretreatment measures in our model-based estimation of the retention effects. Although this procedure will likely improve the precision of treatment effect estimation, measurement errors contained in these pretreatment measures may introduce new bias to the treatment effect estimate. Hence, we estimate the true score of each of these pretreatment measures in Step 3. Then in Step 4, we analyze multivariate multilevel models for estimating the average retention effects on each set of social-emotional outcomes. We include in these models the estimated pretreatment true scores along with indicators for the propensity strata. Finally, in Step 5, we assess the sensitivity of our conclusions to possible influences of unmeasured confounders. We carry out most of our analysis in HLM 6.0 (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2004) and SPSS 15.0 (SPSS, Inc., 2006).

Step 1: Propensity score estimation. In the ECLS-K data, we identify 207 observed pretreatment covariates that are bivariately associated with kindergarten retention in Year 1. Through a stepwise procedure, we specify a propensity model represented as a hierarchical logistic regression model with children at Level 1 and schools at Level 2 and use maximum likelihood to impute missing data in the predictors for the propensity model. For child *i* attending pretreatment school j in Year 0, the child's conditional probability of being retained in Year 1 is a function of a vector of observed pretreatment personal and classroom characteristics $\mathbf{X}_{ij}$, a vector of school characteristics $\mathbf{W}_j$, and the residual random effect of the pretreatment school j, denoted by u_j^*. Combining all the Level 1 and Level 2 equations, we write the model in the following form:

$$\ln\left[\frac{\Pr(Z_{ij}=1|\mathbf{X}_{ij},\mathbf{W}_j,u_j^*)}{1-\Pr(Z_{ij}=1|\mathbf{X}_{ij},\mathbf{W}_j,u_j^*)}\right]$$

$$=\gamma_{00}+\sum_{g=1}^{G}\gamma_{0g}X_{gij}+\sum_{h=1}^{H}\gamma_{h0}W_{hj}+u_j^*. \quad (1)$$

Our propensity model includes 39 predictors and seven quadratic terms. We use an empirical Bayes estimate of u_j^* to capture the impact of unmeasured school-level predictors of retention. See Appendix A in the supplemental material for additional details about this procedure.

Step 2: Propensity score stratification. Using the logit of the estimated propensity score as an index to indicate the extent to which a child was at risk of repeating kindergarten, we compare the distributions of the logit of propensity of retention between the retained group and the promoted group. This enables us to identify 3,087 promoted children who do not have matches in the retained group. For these children, the risk of retention was essentially null. Hence, we restrict our causal inference to the remaining sample of 7,639 children who had a nonzero probability of being retained. In general, a child was more likely to be retained in kindergarten if the child was a boy, relatively young in age, and from a family with low socioeconomic status; if the child had relatively low academic performance; and if the child demonstrated emotional or behavioral problems. In addition, the schools that the retained kindergartners attended during their pretreatment year tended to have inadequate resources. Their kindergarten teachers tended to hold different standards based on children's capabilities and usually spent less time in reading instruction covering lower-level content.

We divide the sample of at-risk children into 15 strata on the basis of the estimated logit of propensity score. Table 2 compares the within-stratum distributions of the logit of propensity between the retained group and the promoted group. The eight retained students in the last stratum had no matches in the promoted group. Within each of the remaining 14 strata, the two treatment groups had similar distributions of the estimated logit of propensity. The result of hypothesis testing controlling for stratum membership shows no statistically significant between-group difference in 97% of the 207 pretreatment covariates. See Appendix B in the supplemental materials for SPSS syntax.

Table 2
Within-Stratum Distribution of the Logit of Propensity Score for Kindergarten Retention

	Retained			Promoted		
Stratum	*n*	*M*	*SD*	*n*	*M*	*SD*
$L = 0$	0			3,087	−7.12	0.84
$L = 1$	9	−5.42	0.40	3,054	−5.38	0.39
$L = 2$	12	−4.25	0.24	1,661	−4.27	0.25
$L = 3$	14	−3.42	0.20	983	−3.50	0.20
$L = 4$	12	−2.98	0.10	323	−2.97	0.10
$L = 5$	24	−2.57	0.17	441	−2.55	0.16
$L = 6$	23	−2.15	0.07	154	−2.14	0.07
$L = 7$	47	−1.75	0.15	213	−1.77	0.15
$L = 8$	48	−1.24	0.14	143	−1.27	0.14
$L = 9$	46	−0.83	0.11	85	−0.86	0.10
$L = 10$	48	−0.46	0.12	49	−0.47	0.11
$L = 11$	47	−0.01	0.11	35	−0.04	0.15
$L = 12$	49	0.53	0.22	16	0.56	0.17
$L = 13$	45	1.16	0.20	8	1.14	0.17
$L = 14$	39	1.99	0.33	3	1.80	0.17
$L = 15$	8	3.70	1.09	0		
Total	471	−0.66	1.75	10,255	−5.07	1.82

Step 3: Pretreatment true score estimation. In classical test theory, each observed score can be represented as a function of its true score and a measurement error, $Y = \pi + e$. The variance of the observed scale score σ^2_Y is simply the sum of the true score variance σ^2_π and the measurement error variance σ^2_e. The reliability (λ) of a scale score represents the ratio of the true score variance to the observed score variance. Hence, we can compute the error variance of each scale score by applying the formula, $\sigma^2_e = (1 - \lambda)\sigma^2_Y$. We estimate the true scores of a pair of pretreatment measures in each domain through analyzing a multivariate three-level model with two measurement models at Level 1, children at Level 2, and schools at Level 3. Take for example the two most important predictors of child self-perceived competence and interest in peer relationships. The Level 1 model contains two measurement models for the pretreatment teacher and parent ratings, respectively, of a child's interpersonal relationships.

$$X_{mij}=D_{X1ij}(\pi_{X1ij}+e_{X1ij})+D_{X2ij}(\pi_{X2ij}+e_{X2ij}),$$
$$e_{X1ij}\sim N(0,\sigma^2_{e.X1}),\quad e_{X2ij}\sim N(0,\sigma^2_{e.X2}). \quad (2)$$

Here X_{mij}, $m = 1, 2$, are the observed scores of pretreatment teacher and parent ratings, respectively, for child i attending school j. The two observed outcomes are linked to the two measurement models through the dummy indicators—D_{X1ij} for teacher ratings and D_{X2ij} for parent ratings; π_{X1ij} and π_{X2ij} are the true scores of teacher and parent ratings, respectively; e_{X1ij} and e_{X2ij} are the corresponding measurement errors. By convention, we assume that these measurement errors are independent of each other for a given child. Using the reliability information in the ECLS-K data, we compute the error variances $\sigma^2_{e.X1}$ and $\sigma^2_{e.X2}$ and include such information as values of a new variable in the Level 1 data file.

We specify the Level 2 model as follows:

$$\pi_{X1ij}=\beta_{X10j}+\beta_{X11j}Z_{ij}+\sum_{s=2}^{15}\beta_{X1sj}L_{sij}+\beta_{X116j}(Logit\_q)_{ij}+r_{X1ij},$$

$$\pi_{X2ij}=\beta_{X20j}+\beta_{X21j}Z_{ij}+\sum_{s=2}^{15}\beta_{X2sj}L_{sij}+\beta_{X216j}(Logit\_q)_{ij}+r_{X2ij};$$

$$\begin{pmatrix} r_{X1ij} \\ r_{X2ij} \end{pmatrix}\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix},\begin{pmatrix} \tau_{\pi X1} & \tau_{\pi X1.X2} \\ \tau_{\pi X2.X1} & \tau_{\pi X2} \end{pmatrix}\right]. \quad (3)$$

Here L_{sij}, $s = 2, \ldots, 15$, are dummy indicators for 14 of the 15 propensity strata that subclassify the at-risk students. We make additional within-stratum adjustments for the logit of propensity, denoted with $(Logit\_q)_{ij}$. The Level 3 model has 34 equations corresponding to the 34 coefficients at Level 2:

$$\beta_{X10j} = \gamma_{X100} + u_{X10j},$$

$$\beta_{X1sj} = \gamma_{X1s0}, \text{ for } s = 1, \ldots, 16$$

$$\beta_{X20j} = \gamma_{X200} + u_{X20j},$$

$$\beta_{X2sj} = \gamma_{X2s0}, \text{ for } s = 1, \ldots, 16.$$

$$\begin{pmatrix} u_{X10j} \\ u_{X20j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{\beta X10} & \tau_{\beta X10.X20} \\ \tau_{\beta X20.X10} & \tau_{\beta X20} \end{pmatrix}\right] \quad (4)$$

This three-level model estimates each pretreatment true score as follows:

$$\pi^*_{Xmij} = \lambda_{Xm}X_{mij} + (1 - \lambda_{Xm})\left[\gamma_{Xm00} + \gamma_{Xm10}Z_{ij} + \sum_{s=2}^{15}\gamma_{Xms0}L_{sij} + \gamma_{Xm160}(Logit\_q)_{ij} + u^*_{Xm0j}\right], \quad (5)$$

where u^*_{Xm0j} is an empirical Bayes estimate of u_{Xm0j}. The estimate of a child's pretreatment true score π^*_{Xmij} is equal to the corresponding observed score X_{mij} if there is no measurement error, that is, if $\lambda_{Xm} = 1.0$. As the reliability of the observed score decreases, this model draws an increasing amount of information from the predicted mean pretreatment true score of the subpopulation as defined by the child's treatment group membership, propensity of retention, and school membership. Appendix C in the supplemental material explains how to organize the Level 1 data file and how to locate the estimated pretreatment true scores in the Level 2 residual file in HLM 6.0.

Step 4: Model-based retention effect estimation. Clustering the outcome measures by domains, we analyze three multivariate, multilevel models in correspondence with the three sets of social-emotional outcomes: (a) child reports of self-perceived competence and interest in reading, mathematics, and all subjects in Years 3 and 5; (b) child reports of self-perceived competence and interest in peer relationships in Years 3 and 5; and (c) parent, teacher, and child reports of internalizing problem behaviors in Years 1, 3, and 5. Below we use models for child reports of self-perceived peer relationships in Years 3 and 5 to illustrate the procedure. The Level 1 model contains the two measurement models for the Year 3 outcome and the Year 5 outcome denoted with Y_{1ij} and Y_{2ij}, respectively.

$$Y_{mij} = D_{Y1ij}(\pi_{Y1ij} + e_{Y1ij}) + D_{Y2ij}(\pi_{Y2ij} + e_{Y2ij}),$$

$$e_{Y1ij} \sim N(0, \sigma^2_{e.Y1}), \quad e_{Y2ij} \sim N(0, \sigma^2_{e.Y2}). \quad (6)$$

The Year 3 and Year 5 true score outcomes π_{Y1ij} and π_{Y2ij} become latent outcomes at Level 2. The estimated pretreatment true scores of teacher and parent ratings of child interpersonal relationships in Year 0, π^*_{X1ij} and π^*_{X2ij}, obtained from Step 3, are included here as additional pretreatment covariates.

$$\pi_{Y1ij} = \beta_{Y10j} + \beta_{Y11j}Z_{ij} + \sum_{s=2}^{15}\beta_{Y1sj}L_{sij} + \beta_{Y116j}(Logit\_q)_{ij} + \beta_{Y117j}\pi^*_{X1ij} + \beta_{Y118j}\pi^*_{X2ij} + r_{Y1ij},$$

$$\pi_{Y2ij} = \beta_{Y20j} + \beta_{Y21j}Z_{ij} + \sum_{s=2}^{15}\beta_{Y2sj}L_{sij} + \beta_{Y216j}(Logit\_q)_{ij} + \beta_{Y217j}\pi^*_{X1ij} + \beta_{Y218j}\pi^*_{X2ij} + r_{Y2ij};.$$

$$\begin{pmatrix} r_{Y1ij} \\ r_{Y2ij} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{\pi Y1} & \tau_{\pi Y1.Y2} \\ \tau_{\pi Y2.Y1} & \tau_{\pi Y2} \end{pmatrix}\right] \quad (7)$$

At Level 3, γ_{Y110} and γ_{Y210} estimate the effects of kindergarten retention versus promotion on children's self-perceived peer relationships in Years 3 and 5, respectively. The fixed part of the Level 3 model will have 40 equations corresponding to the 40 coefficients at Level 2:

$$\beta_{Y10j} = \gamma_{Y100} + u_{Y10j},$$

$$\beta_{Y11j} = \gamma_{Y110} + u_{Y11j},$$

$$\beta_{Y1sj} = \gamma_{Y1s0}, \text{ for } s = 2, \ldots, 18$$

$$\beta_{Y20j} = \gamma_{Y200} + u_{Y20j},$$

$$\beta_{Y21j} = \gamma_{Y210} + u_{Y21j},$$

$$\beta_{Y2sj} = \gamma_{Y2s0}, \text{ for } s = 2, \ldots, 18.$$

$$\begin{pmatrix} u_{Y10j} \\ u_{Y11j} \\ u_{Y20j} \\ u_{Y21j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{\beta Y10} & \tau_{\beta Y10.Y11} & \tau_{\beta Y10.Y20} & \tau_{\beta Y10.Y21} \\ \tau_{\beta Y11.Y10} & \tau_{\beta Y11} & \tau_{\beta Y11.Y20} & \tau_{\beta Y11.Y21} \\ \tau_{\beta Y20.Y10} & \tau_{\beta Y20.Y11} & \tau_{\beta Y20} & \tau_{\beta Y20.Y21} \\ \tau_{\beta Y21.Y10} & \tau_{\beta Y21.Y11} & \tau_{\beta Y21.Y20} & \tau_{\beta Y21} \end{pmatrix}\right] \quad (8)$$

By testing the null hypotheses $\tau_{\beta Y11} = 0$ and $\tau_{\beta Y21} = 0$, we can investigate whether the retention effect on the Year 3 outcome and that on the Year 5 outcome depend on which school a child attended.

Step 5: Sensitivity analysis. The estimates of the kindergarten retention effects will be unbiased under the assumption that the treatment assignment is independent of the unobserved confounders given the observed covariates. We examine the extent to which our causal conclusions would be altered by additional adjustments for potential unmeasured confounders, the omission of which would create a bias comparable to that of the most important observed covariates (Lin, Psaty, & Kronmal 1998; Rosenbaum, 1986, 2002). In our multilevel context, we imagine that there might be a student-level unmeasured composite U_X and, simultaneously, a school-level unmeasured composite U_W. The impact of the omission of U_X and U_W would depend on their respective associations with the outcome, represented by γ_W and γ_X, conditional on treatment and propensity strata, and would also depend on their respective associations with the treatment assignment, represented by $E[U_{W1}] - E[U_{W0}]$ and $E[U_{X1}] - E[U_{X0}]$ (Hong, 2004; Hong & Raudenbush, 2006). Suppose that the original estimate of retention effect $\hat{\gamma}$ is statistically significant. We imagine that the confounding effects of the hypothetical composites U_X and U_W would be as severe as the strongest confounders among the 207 pretreatment covariates. After making additional adjustment for U_X and U_W, we obtain new estimates of the retention effect

$$\hat{\gamma}^* = \hat{\gamma} \pm \{|\gamma_W \times [E(U_{W1}) - E(U_{W0})]| + |\gamma_X \times [E(U_{X1}) - E(U_{X0})]|\} \quad (9)$$

If the confidence interval for $\hat{\gamma}^*$ does not contain zero, we will consider our original conclusion about the retention effect to be insensitive to the impact of unmeasured confounders.

Results

Distributions of True Scores at the Child Level and the School Level

Child self-perceived competence and interest in academic learning. We found that, on average, child self-perceived competence and interest in reading, math, and all subjects dropped significantly from Year 3 to Year 5 (see Table 3). The estimated cross-year mean differences were –0.28 in reading, $\chi^2(1) = 594.88$, $p < .001$, −0.25 in mathematics, $\chi^2(1) = 383.43$, $p < .001$, and –0.20 in all subjects, $\chi^2(1) = 338.25$, $p < .001$. Only about 2%–7% of the true score variance in each of these outcomes existed between schools in Years 3 and 5, which indicates a great amount of within-school variation among children in these measures. This is perhaps because children formed their self-perceptions largely through evaluating their own relative academic standings among the peers in the same school. Interestingly, at the child level, self-rated competence and interest in reading and that in mathematics were not highly correlated ($r = .17$; see Table 4) in Year 3, though competence and interest in all subjects were correlated with both competence and interest in reading ($r = .61$) and in math ($r = .62$). We observed a similar pattern for the corresponding Year 5 measures.

Child self-perceived competence and interest in peer relationships. There was a slight but significant decrease in the average self-perceived competence and interest in peer relationships from Year 3 to Year 5 (see Table 5). The estimated cross-year difference was –0.05, $\chi^2(1) = 26.70$, $p < .001$. About 3%–5% of the true score variance in each year lay between schools. Again, we reason that children perceived their own popularity in comparison with that of their peers in the same school. The correlation between the Year 3 and Year 5 measures was .52 at both the child level and the school level.

Table 3
Distributions of True Scores Measuring Children's Social-Emotional Development: Children's Self-Perceived Competence and Interest in Academic Learning

Subject area and year	M	Cross-year mean diff.	$\chi^2(1, N = 7{,}612)$	Child-level variance	School-level variance
Reading		−0.28	594.88**		
Year 3	3.24			0.36	0.02
Year 5	2.96			0.46	0.03
Math		−0.25	383.43**		
Year 3	3.13			0.54	0.03
Year 5	2.88			0.56	0.01
All subjects		−0.20	338.25**		
Year 3	2.89			0.31	0.02
Year 5	2.68			0.32	0.02

** $p < .001$.

Internalizing problem behaviors. On average, teacher ratings of child internalizing problem behaviors showed a small but significant increase from Year 1 to Year 3 (mean difference = 0.04), $\chi^2(1) = 14.14$, $p < .001$, and appeared to have no change from Year 3 to Year 5 (mean difference = 0.01), $\chi^2(1) = 0.58$, $p > .05$. Meanwhile, the average child self-rating of internalizing problem behaviors decreased by a considerable amount from Year 3 to Year 5 (mean difference = −0.16), $\chi^2(1) = 269.57$, $p < .001$ (see Table 6). We found about 11%–13% of the true score variance existing between schools in internalizing problem behaviors as rated by teachers, parents, and children themselves in Years 1, 3, and 5. Teacher ratings across the three time points were correlated at .29–.42 at the child level and .15–.28 at the school level (see Table 7). Parent ratings showed a weak correlation with teacher ratings in Year 1 both at the child level ($r = .25$) and at the school level ($r = .17$). Child self-ratings were only weakly correlated with teacher ratings and parent ratings, indicating a certain degree of disagreement among the raters. Child self-ratings in Year 1 and Year 3 were correlated at .55 at the child level and as high as .97 at the school level, the latter suggesting a consistent and strong school impact on children's self-perceptions of internalizing problem behaviors.

Causal Effects of Kindergarten Retention

Prior to the treatment year, the retained children were behind the same-age promoted children on average in most of the cognitive and social-emotional domains. As we described earlier, the two groups also differed in a large number of pretreatment covariates, including child demographic characteristics, family backgrounds, and prior learning experiences at home and in school. Hence, a direct comparison of the posttreatment social-emotional outcomes between the retained children and the promoted children would not tell us how the retained children would have performed had they been promoted to the first grade instead. Following the five steps laid out in the previous section on analytic procedure, we estimated the effects of kindergarten retention on the retained students' self-perceived competence and interest in academic learning, peer relationships, and internalizing problem behaviors at the end of the retention year, 2 years later, and 4 years later. In examining the mean differences between the retained group and the promoted group across the 15 propensity strata, we observed no systematic association between a child's propensity of repeating kindergarten and the retention effect on each of these outcome measures. Below we report the model-based estimation of the retention effects (see Table 8).

Child self-perceived competence and interest in academic learning. In comparison with the promoted children at similar risk of repetition, the retained students reported a higher level of self-perceived competence and interest in reading in Year 3 (coefficient = 0.12, standard error = 0.05, $t = 2.49$) and Year 5 (coefficient = 0.10, standard error = 0.05, $t = 1.91$), although the estimate for the Year 5 outcome failed to reach a significance level

Table 4
Correlations of Children's Ratings of Competence and Interest in Academic Learning

Measure	1	2	3	4	5	6
1. Child self-rating on reading in Year 3	—	.63	.44	.40	.71	.48
2. Child self-rating on reading in Year 5	.50	—	.23	.29	.23	.61
3. Child self-rating on math in Year 3	.17	.00	—	.57	.88	.35
4. Child self-rating on math in Year 5	.08	.19	.44	—	.61	.76
5. Child self-rating on all subjects in Year 3	.61	.27	.62	.31	—	.54
6. Child self-rating on all subjects in Year 5	.32	.63	.26	.65	.47	—

Note. The lower triangular matrix shows true score correlations at the child level; the upper triangular matrix shows true score correlations at the school level.

of .05. Using the respective standard deviations of the retained students' self-perceived competence and interest in reading in Year 3 and Year 5 as the basis, we found the effect sizes of the above two estimates to be 0.17 (95% confidence interval = 0.04, 0.30) and 0.14 (95% confidence interval = −0.00, 0.29), respectively. The retained students' self-perceived competence and interest in math, though higher than that of the comparable promoted children in both Year 3 (coefficient = 0.08, standard error = 0.05, t = 1.57, effect size = 0.11, 95% confidence interval = −0.03, 0.24) and Year 5 (coefficient = 0.07, standard error = 0.06, t = 1.23, effect size = 0.09, 95% confidence interval = −0.05, 0.23), did not reach a significance level of .05. In reporting on competence and interest in all school subjects, the retained students' self-ratings were higher than those of the comparable promoted children in Year 3 (coefficient = 0.10, standard error = 0.04, t = 2.29, effect size = 0.16, 95% confidence interval = 0.02, 0.29), but not in Year 5 (coefficient = 0.00, standard error = 0.05, t = 0.01, effect size = 0.00, 95% confidence interval = −0.16, 0.16). According to the results of multivariate hypothesis testing, the retention effects in each of these three subject areas were generally stable from Year 3 to Year 5. The average effect size is about 0.11 over these six measures of self-perceived competence and interest in academic learning. Except for the outcome measure of all subjects in Year 5, the confidence intervals for the retention effects on the other five outcome measures were mostly positive. An omnibus test of the null effects of retention on all the six outcomes shows that the effect of kindergarten retention was significant on at least one of these outcomes, $\chi^2(6) = 14.41$, $p < .05$.

Child self-perceived competence and interest in peer relationships. On average, the retained kindergartners rated themselves higher than did the promoted children at similar risk of repetition in both Year 3 (coefficient = 0.06, standard error = 0.04, t = 1.35, effect size = 0.09, 95% confidence interval = −0.04, 0.23) and Year 5 (coefficient = 0.03, standard error = 0.05, t = 0.56, effect size = 0.05, 95% confidence interval = −0.11, 0.19). However, these estimates were not significantly different from zero.

Table 5
Distributions of True Scores Measuring Children's Social-Emotional Development: Children's Self-Perceived Competence and Interest in Peer Relationships

Year	M	Cross-year mean diff.	$\chi^2(1, N = 7{,}612)$	Child-level variance	School-level variance
3	3.02	−0.05	26.70**	0.31	0.01
5	2.97			0.29	0.02

** $p < .001$.

Internalizing problem behaviors. Based on the teacher ratings, at the end of the treatment year, the retained students showed a lower level of internalizing problem behaviors than did the promoted children at similar risk of repetition (coefficient = −0.08, standard error = 0.04, t = −2.19, effect size = −0.16, 95% confidence interval = −0.30, −0.02). Although the retained students continued to receive lower ratings from their teachers in internalizing problem behaviors 2 years after the retention (coefficient = −0.05, standard error = 0.04, t = −1.07, effect size = −0.08, 95% confidence interval = −0.22, 0.06), the average difference between the retained group and the promoted group was no longer significant. Kindergarten retention showed no effect on the teacher-rated internalizing problem behaviors 4 years after the retention (coefficient = −0.00, standard error = −0.06, t = −0.08, effect size = −0.01, 95% confidence interval = −0.18, 0.16). According to the parent ratings, however, the retained students displayed slightly more internalizing problem behaviors than did the promoted children at similar risk of repetition at the end of the treatment year (coefficient = 0.01, standard error = 0.02, t = 0.41, effect size = 0.02, 95% confidence interval = −0.08, 0.13), though the estimate was not significantly different from zero. A multivariate hypothesis test showed a significant difference between Year 1 teacher ratings and parent ratings in the estimated retention effects on child internalizing problem behaviors, $\chi^2(1) = 4.30$, $p < .05$, results not tabulated. Consistent with the teacher observations, the retained students themselves reported a lower level of internalizing problem behaviors than did the comparable promoted children in both Year 3 (coefficient = −0.11, standard error = 0.05, t = −2.25, effect size = −0.14, 95% confidence interval = −0.26, −0.02) and Year 5 (coefficient = −0.08, standard error = 0.05, t = −1.63, effect size = −0.11, 95% confidence interval = −0.25, 0.02). Although the t statistic for the latter did not reach the significance level of .05, a test of the contrast showed no significant difference in the retention effect estimates between Year 3 and Year 5, $\chi^2(1) = 0.34$, $p > .05$, results not tabulated. Nor did the retention effect estimates differ between teacher ratings and child ratings in either Year 3, $\chi^2(1) = 0.95$, $p > .05$, results not tabulated, or Year 5, $\chi^2(1) = 1.08$, $p > .05$, results not tabulated. The average effect size is about –0.08 over the six measures of internalizing problem behaviors. With parent ratings as the only exception, the confidence intervals for the retention effects on the other five outcomes were mostly

Table 6
Distributions of True Scores Measuring Children's Social-Emotional Development: Internalizing Problem Behaviors

Respondent and year	M	Cross-year mean diff.	$\chi^2(1, N = 7{,}632)$	Child-level variance	School-level variance
Teacher					
Year 1	1.63	0.04	14.14**	0.19	0.02
Year 3	1.67	0.05	19.41**	0.19	0.03
Year 5	1.68	0.01	0.58	0.21	0.03
Parent					
Year 1	1.55			0.09	0.00
Child					
Year 3	2.23	−0.16	269.57**	0.37	0.06
Year 5	2.06			0.27	0.03

** $p < .001$.

negative. Finally, an omnibus test suggested that, under the null hypothesis, we have a fairly small chance of obtaining a significant estimate of the retention effect on at least one of the six outcome measures of internalizing problem behaviors, $\chi^2(6) = 12.51$, $p = .051$.

Results of a sensitivity analysis suggested that our causal conclusions can potentially be altered by hypothetical unmeasured pretreatment covariates that display as strong confounding effects as the most important covariates observed. Although the omission of such confounders is not highly plausible given the richness of the data set, the above-estimated results remain tentative.

Discussion

In this study, we examined the effects of kindergarten retention relative to promotion to the first grade on children's self-perceived competence and interest in academic learning and in peer relationships and the effects on child internalizing problem behaviors as rated by teachers, parents, and children themselves in the early, middle, and later elementary years. Here we summarize the empirical findings, discuss their theoretical implications, review our methodology, and suggest strategies for future research.

The results showed that, 2 years after retention, the retained kindergartners perceived a higher level of competence and interest in academic learning than they would have if they had been promoted to the first grade instead. This is most likely to be true in reading and in all subjects. Retention did not show detectable effects on children's self-perceived competence and interest in peer relationships 2 and 4 years after the treatment. Yet according to teachers' observations at the end of the treatment year and children's self-reports 2 years later, the retained students experienced a lower level of internalizing problem behaviors on average as a result of retention than they would have if promoted. In general, this study has shown no evidence suggesting that kindergarten retention does harm to children's social-emotional development. Rather, it seems that retaining the at-risk children in kindergarten would likely raise their self-confidence and interest, especially in reading and in all subjects, and might even decrease their internalizing problem behaviors. We note that the estimated effect sizes on all the outcome measures are relatively small. Therefore, our results do not indicate that kindergarten retention will bring great benefits to the social-emotional development of all the children who would possibly be retained.

Table 7
Correlations of Teachers', Parents', and Children's Ratings of Internalizing Problem Behaviors

Measure	1	2	3	4	5	6
1. Teacher rating in Year 1	—	.15	.20	.17	.15	.25
2. Teacher rating in Year 3	.34	—	.28	.31	.34	.34
3. Teacher rating in Year 5	.29	.42	—	.01	.43	.38
4. Parent rating in Year 1	.25	.19	.17	—	.21	.30
5. Child self-rating in Year 3	.19	.22	.17	.10	—	.97
6. Child self-rating in Year 5	.18	.18	.21	.14	.55	—

Note. The lower triangular matrix shows true score correlations at the child level; the upper triangular matrix shows true score correlations at the school level.

The early intervention theory and the social comparison theory provide complementary interpretations of the above findings. As hypothesized by the early intervention theory, a second chance of learning the kindergarten curriculum accompanied by the growing maturity in cognition, emotions, and social behaviors may have provided the retained students with a better preparation for learning the academic content in the later years, and therefore may have improved their academic standings among a group of younger peers. These successful experiences may have in turn increased the retained students' self-confidence and fostered their academic interest, as suggested by the social comparison theory. Had these children been promoted to the first grade instead, many of them would have continued to struggle with the content materials and perhaps would have received lower grades than their same-age classmates, which might have led to self-perceptions of lower competence and reduced interest in school subjects as shown in our analytic results. These findings coincide with Hong and Yu's (2007) report that, even though the retained kindergartners displayed a lower level of reading and math knowledge and skills on average at the end of the retention year when compared with the same-age promoted children who had been at similar risk of repetition, the retained students showed a faster growth rate in reading and math and were able to gradually catch up with the comparison group in the later elementary years. Future research needs to investigate whether the retention effects on children's cognitive growth may have been mediated by the retained stu-

Table 8
Model-Based Estimation of Kindergarten Retention Effects on Social-Emotional Development

Domain	Respondent	Year	Coefficient (SE)	ES	95% CI of ES	χ^2 statistic
Perceived interest and competence in academic subjects						$\chi^2(6, N=7{,}612) = 14.41$
Reading	Child	3	0.12*(0.05)	0.17	0.04, 0.31	
		5	0.10 (0.05)	0.14	−0.00, 0.29	
Math		3	0.08 (0.05)	0.11	−0.03, 0.24	
		5	0.07 (0.06)	0.09	−0.05, 0.23	
All subjects		3	0.10*(0.04)	0.16	0.02, 0.29	
		5	0.00 (0.05)	0.00	−0.16, 0.16	
Perceived interest and competence in peer relationships	Child	3	0.06 (0.04)	0.09	−0.04, 0.23	$\chi^2(2, N=7{,}612) = 1.79$
		5	0.03 (0.05)	0.05	−0.11, 0.19	
Internalizing problem behavior	Teacher	1	−0.08*(0.04)	−0.16	−0.30, −0.02	$\chi^2(6, N=7{,}632) = 12.51$
		3	−0.05 (0.04)	−0.08	−0.22, 0.06	
		5	−0.00 (0.06)	−0.01	−0.18, 0.16	
	Parent	1	0.01 (0.02)	0.02	−0.08, 0.13	
	Child	3	−0.11*(0.05)	−0.14	−0.26, −0.02	
		5	−0.08 (0.05)	−0.11	−0.25, 0.02	

Note. ES = effect size; CI = confidence interval.
* $p < .05$.

dents' improved basic knowledge and skills, by their enhanced self-confidence and interest in school subjects, or by both.

The results from the current study contradict the predictions of the labeling theory. Being retained in kindergarten did not seem to have alienated the retained students from their new peer groups. Nor did the retained students develop more negative feelings about themselves on average than they would have if promoted. We reason that, even if there was stigma associated with retention in general, the new kindergartners, who were yet to be socialized in the school setting, were unlikely to create a strong negative opinion environment for the retained students. This may explain why the retained students seemed to have no more difficulties in forming friendships than did their at-risk promoted peers. On the contrary, as we can infer from the teachers' observations, when children at risk of repetition were promoted to the first grade, being a "low achiever" and struggling at the bottom of the first-grade class may have intensified feelings of anxiety, shame because of failure, and even depression, causing relatively more internalizing problem behaviors among these children than among their retained counterparts at the end of the retention year. Furthermore, some of the promoted at-risk children might be retained in the later years. If retention at a higher grade level stigmatized these children to a greater extent than did early retention, as suggested by the previous literature (Finlayson, 1977; Morrison & Perry, 1956), while the stigma, if there was any, associated with kindergarten retention might have been washed away by time in the later years, this would provide a possible explanation for why the retained kindergartners perceived a lower level of internalizing problem behaviors 2 years after retention in comparison with their same-age promoted peers who were similarly at risk.

We have adapted statistical methods appropriate for addressing our causal questions at hand. In summary, we used propensity score stratification to remove selection bias associated with a very large number of observed covariates. By comparing the propensity score distributions between the retained group and the promoted group, we empirically defined a population of children at risk for repeating kindergarten. This is the target population of children to which we generalize our causal inference about the effects of kindergarten retention. We accounted for measurement errors in the key predictors as well as in the outcomes. In addition, we made adjustments for dependence among multiple observations per child and among multiple children per school through analyzing a series of multivariate, multilevel models. Below we discuss conditions under which these methods are appropriate for use.

Strong Ignorability Assumption

In general, propensity score adjustment methods should generate unbiased estimates of treatment effects from nonexperimental data under the strong ignorability assumption. That is, given all the observed pretreatment covariates, treatment assignment is independent of all the unmeasured confounders. The ECLS-K data set contains very rich pretreatment information about the students, their families, teachers, and schools. Our list of observed pretreatment covariates includes all the factors typically considered by educators and parents in making decisions about grade retention (Grant & Richardson, 1998). Among them, the most important predictors of kindergarten retention and of children's social-emotional outcomes are pretreatment measures in various cognitive and social-emotional domains. Therefore, we have reason to assume that, once we have adjusted for these observed pretreatment covariates, the estimation of the retention effects may not be severely confounded by unmeasured covariates. Even though it is unlikely that we have omitted many potentially important confounders, we conducted a sensitivity analysis to test the robustness of our results in the presence of hypothetical child-level and school-level confounders comparable to the most important observed confounders. The retention effects that we estimated to be statistically significant were all relatively small in magnitude. For this reason, once we made further adjustments for the hypothetical confounders, the new estimates could no longer be distinguished from zero. Hence, our conclusion about the positive effects of kindergarten retention on children's social-emotional outcomes was potentially sensitive to the influence of unmeasured confounders.

Variable Selection for the Propensity Score Model

There are several different approaches to selecting variables for the propensity score model. Suppose that Set A includes all the predictors of the treatment, and Set B includes all the predictors of a particular outcome. Some researchers have recommended choosing variables from Set B for the propensity score model in the interest of removing selection bias and in the meantime improving the precision of the treatment effect estimate (Brookhart et al., 2006). Although preferable when the analysis involves a single outcome, this method becomes cumbersome in practice when there are multiple outcomes under consideration. In our case, we examined the kindergarten retention effects on as many as 14 different outcomes in a variety of social-emotional domains. We therefore opted for selecting variables from Set A through a stepwise procedure in logistic regression. By adopting this approach, we were able to apply a single propensity score model to the 14 different outcomes. We then stratified our analytic sample such that about 97% of more than 200 predictors of kindergarten retention show no significant within-stratum differences between the retained group and the promoted group. Under strong ignorability, our approach to propensity score stratification approximates a randomized block design. That is, within each stratum, the retained kindergartners and their promoted counterparts differ only in their treatment assignment. Hence, systematic differences in the observed outcomes between these two groups are to be attributed to the treatments only. Previous research has suggested that combining propensity score matching or stratification with covariance adjustment for prognostic predictors of the outcome effectively improves the precision of treatment effect estimation (Rubin & Thomas, 2000). Because many predictors of kindergarten retention also predict children's social-emotional outcomes, and because we have made additional adjustments for a pair of most important predictors for each set of outcomes, we reason that selecting variables from Set A should not have cost us precision in estimating the retention effects.

Accounting for Measurement Errors

As we have explained earlier, measurement errors in predictors are typically more consequential than measurement errors in outcomes, because the former may bias the treatment effect estimate when the treatment is associated with the error-laden predictors. To remove measurement errors in the observed pretest scores, we obtained the empirical Bayes estimates of their corresponding true scores. Entering these estimated true scores instead of using the observed scores as covariates in the outcome models provided a safeguard against additional bias introduced by the measurement errors in these predictors. We acknowledge that there might be measurement errors in some other predictors selected for the propensity model. Further adjustment would have been possible had information about the psychometric properties of those other predictors become available. Accounting for the measurement errors in the outcomes has its unique value in multivariate, multilevel analysis. This is because, by using the computed error variances at Level 1, we were able to specify a child-specific random coefficient at Level 2 for each outcome, thereby separating the error variance from the true score variance among children within each school as well as from the true score variance between schools.

Heterogeneous Retention Effects

Because kindergarten retention was administered by school organizations, the retention effects may depend on which school a child attended. The school-specific retention effects on the social-emotional outcomes, if correlated across the different outcomes, may suggest a need for investigating various treatment settings that have made kindergarten retention more or less effective. Due to the small number of retained kindergartners sampled in each school, we were unable to investigate in this study the school-level variation and covariation of the retention effects. Future studies with more sampled children in both treatment groups per school will enable evaluations of the effectiveness of alternative school practices designed for helping the retained children and the at-risk promoted children. The retention effects may also depend on child characteristics and may differ among children who were retained for different reasons. For example, some children were retained in kindergarten due to behavioral problems; others were retained because they fell behind in academic subjects; some others who had displayed neither behavioral problems nor academic difficulties were nonetheless retained, often because their parents expected them to gain relative advantage among younger peers. To study whether the retention effects were differential across these subgroups of children, one could include in the outcome model interactions between the treatment indicator and indicators for different reasons of retention.

Retention Effects for the Population of Retained Students

We have intended to generalize our estimation results to all children at risk of kindergarten retention as represented by our analytic sample. As an alternative, researchers may choose a population of retained students as the target population. To do so, we would weight the retention effect estimate within each propensity stratum by the proportion of retained students allocated to that stratum. The weighted sample would then represent a population of retained kindergartners. We can compute the weighted average retention effects across all the strata. The results would provide information about the expected counterfactual outcomes associated with promotion for children who were actually retained.

References

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

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163*(12), 1149–1156.

Byrnes, D. A. (1989). Attitudes of students, parents, and educators toward repeating a grade. In L. A. Shepard & M. L. Smith (Eds.), *Flunking grades: Research and policies on retention* (pp. 108–131). New York: Falmer Press.

Cheong, Y. F., & Raudenbush, S. W. (2000). Measurement and structural models for children's problem behaviors. *Psychological Methods, 5*(4), 477–495.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24*(2), 295–313.

Dawson, P. (1998). A primer on student retention: What the research says. *Communique* (Milwaukee, WI), *26*(8), 28–30.

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

Finlayson, H. J. (1977). Nonpromotion and self-concept development. *Phi Delta Kappan, 59,* 205–206.

Grant, J., & Richardson, I. (1998). *The retention/promotion checklist.* Peterborough, NH: Crystal Springs Books.

Gresham, F., & Elliot, S. (1990). *Social skills rating system.* Circle Pines, MN: American Guidance Services.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hill, J. L., Waldfogel, J., Brooks-Gunn, J., & Han, W.-J. (2005). Maternal employment and child development: A fresh look using newer methods. *Developmental Psychology, 41*(6), 833–850.

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

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

Hong, G. (2004). *Causal inference for multi-level observational data with application to kindergarten retention.* Unpublished doctoral dissertation, University of Michigan, Ann Arbor.

Hong, G. (2007, January). *Marginal mean weighting adjustment for selection bias.* Invited presentation at the University of Chicago Education Workshop, Chicago, IL.

Hong, G., & Hong, Y. (2007, December). *Reading instruction time and homogeneous grouping in kindergarten: An application of the marginal mean weighting method.* Invited presentation at the University of Michigan, School of Education, Ann Arbor, MI.

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.

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.

Huang, I.-C., Frangakis, C., Dominici, F., Diette, G. B., & Wu, A. W. (2005). Approach for risk adjustment in profiling multiple physician groups on asthma care. *Health Services Research, 40,* 253–278.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika, 87,* 706–710.

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

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

Lemert, E. M. (1967). *Human deviance, social problems, and social control.* Englewood Cliffs, NJ: Prentice-Hall.

Lin, D. Y., Psaty, B. M., & Kronmal, R. A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics, 54,* 948–963.

Little, R. J. A. (1985). A note about models for selectivity bias. *Econometrica, 53*(6), 1469–1474.

Little, R. J., An, H., & Johanns, J. (2000). A comparison of subset selection and analysis of covariance for the adjustment of confounders. *Psychological Methods, 5*(4), 459–476.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68,* 304–305.

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

Marsh, H. (1990). *Self-Description Questionnaire manual.* Campbelltown, New South Wales, Australia: University of Western Sydney, Macarthur.

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

Morrison, F. J., Griffith, E. M., & Alberts, D. M. (1997). Nature–nurture in the classroom: Entrance age, school readiness, and learning in children. *Developmental Psychology, 33*(2), 254–262.

Morrison, I. E., & Perry, I. F. (1956). Acceptance of overage children by classmates. *Elementary School Journal, 56,* 217–220.

National Center for Education Statistics. (2002). *Early Childhood Longitudinal Study—Kindergarten Class of 1998–99 (ECLS-K) psychometric report for kindergarten through first grade* (Working Paper No. 200205). Washington, DC: Author.

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

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

Plummer, D. L., & Graziano, W. G. (1987). Impact of grade retention on the social development of elementary school children. *Developmental Psychology, 23*(2), 267–275.

Pollack, J. M., Atkins-Burnett, S., Tourangeau, K., & West, J. (2005). *Early Childhood Longitudinal Study—Kindergarten Class of 1998–99 (ECLS-K) psychometric report for the third grade* (Tech. Rep. No. 2005062). Washington, DC: National Center for Education Statistics.

Pollack, J. M., Najarian, M., Rock, D. A., Atkins-Burnett, S., & Hausken, E. G. (2005). *Early Childhood Longitudinal Study—Kindergarten Class of 1998–99 (ECLS-K) psychometric report for the fifth grade* (Tech. Rep. No. 2006036rev). Washington, DC: National Center for Education Statistics.

Quay, H. C., & Peterson, D. (1987). *Manual for the Revised Behavior Problem Checklist.* Coral Gables, FL: University of Miami.

Raudenbush, S., Bryk, A., Cheong, Y. F., Congdon, R., & du Toit, M. (2004), *HLM6: Hierarchical linear and nonlinear modeling.* Lincolnwood, IL: Scientific Software International.

Raudenbush, S. W., Brennan, R. T., & Barnett, R. C. (1995). A multivariate hierarchical model for studying psychological change within married couples. *Journal of Family Psychology, 9*(2), 161–174.

Raudenbush, S. W., & Bryk, A. S. (2002), *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model for studying school climate in secondary schools with estimation via the EM algorithm. *Journal of Educational Statistics, 16*(4), 295–330.

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

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In M. E. Halloran and D. Berry (Eds.), *Statistical models in epidemiology, the environment, and clinical trials* (pp. 95–134). New York: Springer.

Robins, J. M., Hernan, M., & Siebert, U. (2003). Effects of multiple interventions. *Population Health Metrics, 2,* 2191–2230.

Rosenbaum, P. R. (1984). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association, 79*(385), 41–48.

Rosenbaum, P. R. (1986). Dropping out of high school in the United States: An observational study. *Journal of Educational Statistics, 11*, 207–224.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association, 82*(398), 387–394.

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

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.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516–524.

Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association, 81,* 961–962.

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

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

Shepard, L. A. (1989). A review of research on kindergarten retention. In L. A. Shepard and M. L. Smith (Eds.), *Flunking grades: Research and policies on retention* (pp. 64–78). Philadelphia: Palmer Press.

Shepard, L. A., & Smith, M. L. (1989). Academic and emotional effects of kindergarten retention in one school district. In L. A. Shepard & M. L. Smith (Eds.), *Flunking grades: Research and policies on retention* (pp. 79–107). London: Falmer Press.

SPSS, Inc. (2006). *SPSS 15.0 base user's guide.* Upper Saddle River, NJ: Prentice Hall.

Stone, R. (1993). The assumption on which causal inferences rest. *Journal of the Royal Statistical Society, Series B (Methodological), 55*(2), 455–466.

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

U.S. Bureau of the Census. (1995). *Current population survey.* Washington, DC: Author.

Zill, N., Loomis, L. S., & West, J. (1997). *The elementary school performance and adjustment of children who enter kindergarten late or repeat kindergarten: Findings from national surveys* (Stat. Anal. Rep. NCES 98–097). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Received December 19, 2006
Revision received July 18, 2007
Accepted December 20, 2007 ■