



Teacher–child relationships and academic achievement: A multilevel propensity score model approach [☆]

Meghan P. McCormick ^{a,*}, Erin E. O'Connor ^b, Elise Cappella ^a, Sandee G. McClowry ^a

^a Department of Applied Psychology, New York University, USA

^b Department of Teaching and Learning, New York University, USA

ARTICLE INFO

Article history:

Received 7 September 2012

Received in revised form 26 April 2013

Accepted 1 May 2013

Keywords:

Teacher–child relationships

Achievement

Methodology

Causal inference

ABSTRACT

A robust body of research finds positive cross-sectional and longitudinal associations between teacher–child relationships and children's academic achievement in elementary school. Estimating the causal effect of teacher–child relationships on children's academic achievement, however, is challenged by selection bias at the individual and school level. To address these issues, we used two multilevel propensity score matching approaches to estimate the effect of high-quality teacher–child relationships in kindergarten on math and reading achievement during children's transition to first grade. Multi-informant data were collected on 324 low-income, Black and Hispanic students, and 112 kindergarten and first-grade teachers. Results revealed significant effects of high-quality teacher–child relationships in kindergarten on math achievement in first grade. No significant effects of teacher–child relationships were detected for reading achievement. Implications for intervention development and public policy are discussed.

© 2013 Society for the Study of School Psychology. Published by Elsevier Ltd. All rights reserved.

1. Introduction

A robust body of research has identified associations between high-quality teacher–child relationships—characterized by high levels of closeness and low levels of conflict—and children's academic achievement in elementary school (Birch & Ladd, 1997; Hamre & Pianta, 2001; O'Connor & McCartney, 2007; Pianta & Stuhlman, 2004; Rudasill, 2011). Additional studies find that high-quality teacher–child relationships may promote academic resilience among lower-income, racial/ethnic minority children at-risk for poor achievement (Crosnoe et al., 2010; Murray & Zvoch, 2011). This formative work suggests that interventions designed to boost academic achievement in lower-income urban schools should consider targeting teacher–child relationship quality. Research, however, has yet to use multilevel models to infer causal impacts of high-quality teacher–child relationships on academic achievement within this high-risk population of students and schools. To address this need, we used multilevel propensity score models to estimate the effects of high-quality teacher–child relationships in kindergarten on standardized measures of student math and reading achievement in first grade in 22 urban elementary schools. We hypothesized significant effects of high-quality teacher–child relationships on math and reading achievement.

1.1. Teacher–child relationships and academic achievement during the transition to school

Teacher–child relationships are bidirectional, interpersonal exchanges that take place in proximal (e.g., the interpersonal interaction) and distal systems (e.g., the classroom context) (Bronfenbrenner & Morris, 1998; Pianta, 1999). Conceptual studies,

[☆] The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through grants R305B080019 and R305A080512 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.

* Corresponding author at: New York University, Applied Psychology Department, 246 Greene Street, New York, NY 10003, USA.

E-mail address: meghan.mccormick@nyu.edu (M.P. McCormick).

ACTION EDITOR: Scott Graves.

based in attachment theory, propose that children who experience these high-quality relationships are able to rely on teachers as a secure base and a resource for actively exploring the school environment (Howes, Phillipsen, & Peisner-Feinberg, 2000; Hughes, Cavell, & Wilson, 2001). Thus, high-quality teacher–child relationships may boost students' learning by creating a supportive environment in which children are motivated to actively and appropriately engage in the classroom (Ladd & Burgess, 1999).

Recent studies have found that the protective effect of teacher–child relationships on academic achievement may be stronger for lower-income and racial/ethnic minority students, compared to more affluent, White students (Maldonado-Carreño & Votruba-Drzal, 2011; Wu, Hughes, & Kwok, 2010). However, children of lower socioeconomic status tend to be at higher-risk for low-quality relationships with teachers (Pianta & Stuhlman, 2004). In addition, past research has found that White children are likely to have closer relationships with teachers than Black children (Ladd, Birch, & Buhs, 1999). Improving urban schools is of special interest to policymakers interested in shifting resources to close academic achievement gaps (Jacob & Ludwig, 2009). As such, it may be important for future studies of teacher–child relationships to focus attention on lower-income urban schools and students (Jacob & Ludwig, 2009; Murray & Zvoch, 2011).

Theorists argue the transition to elementary school marks a key period for children's development and subsequent achievement (Alexander, Entwisle, Blyth, & McAdoo, 1988). Indeed, early formal schooling experiences are influential in predicting children's highly stable achievement trajectories across childhood and adolescence (Entwisle, Alexander, & Olson, 2005). The transition from kindergarten to first grade appears to be a particularly critical developmental stage for children, due to growing emphasis on emerging literacy and numeracy skills and higher academic expectations (Alexander et al., 1988; Entwisle et al., 2005). This transition is likely to be especially important for children attending lower-income urban schools, as children who enter school with high levels of socioeconomic risk may experience less optimal relationships with their teachers (Jerome, Hamre, & Pianta, 2009).

1.2. Inferring causality between teacher–child relationships and academic achievement

Although the research base linking teacher–child relationships and academic achievement is quite robust (e.g., Roorda, Koomen, Spilt, & Oort, 2011), studies seeking to identify causal effects of teacher–child relationships are limited. Much of the research examining teacher–child relationships and achievement has been nonexperimental (Maldonado-Carreño & Votruba-Drzal, 2011). In general, a central issue in nonexperimental studies is the identification of comparable individuals (e.g., students) to remove selection bias.

Existing studies typically use a number of demographic variables to control for between-child differences that may influence selection into high-quality teacher–child relationships and subsequent academic achievement (e.g., Hughes, 2011; Hughes, Luo, Kwok, & Loyd, 2008; Ladd et al., 1999). Regression analysis attempts to address selection bias by including potential confounding covariates, theoretically and empirically associated with the outcome, in a linear model. However, for regression models to yield causal estimates, they must include all confounding covariates and must be specified correctly. In practice, regression methods that require linearity and additivity may not be appropriate when the model includes a large number of covariates. Because they use prediction equations, regression models extrapolate over portions of the covariate space where there are no data (Gelman & Hill, 2007; Hill, 2011; Hill, Waldfogel, Brooks-Gunn, & Han, 2005). As such, regression models may over or underestimate effects by making comparisons in sections of the covariate space where there is no clear counterfactual for either group.

In addition, although controlling for confounding covariates in a regression is a good first step in limiting selection bias in studies of teacher–child relationships, it is possible that previous analyses omitted a number of important characteristics, such as child sociability, behavior, and intelligence, likely related to both teacher–child relationships and academic achievement. Complicating interpretation is the fact that relations between teacher–child relationships and achievement may actually reflect rater effects if teacher-reported measures were collected (Maldonado-Carreño & Votruba-Drzal, 2011). For example, because teachers are more likely to have high-quality relationships with children who are behaviorally regulated, they may perceive those children to have higher levels of academic skills than less behaviorally regulated children (O'Connor & McCartney, 2007; Rudasill, Reio, Stipanovic, & Taylor, 2010).

Recent studies on teacher–child relationships have begun to address the issue of confounding factors (Ly, Zhou, Chu, & Chen, 2012; Spilt, Hughes, Wu, & Kwok, 2012). Notably, controlling for initial levels of achievement, Maldonado-Carreño and Votruba-Drzal (2011) examined within-child associations in teacher–child relationships and achievement to limit the threat of selection bias and compare effects from teacher-reported and standardized achievement outcomes. Using data from the NICHD Study of Early Child Care and Youth Development from kindergarten to fifth grade, they found no significant associations between teacher–child relationship quality in kindergarten and standardized achievement scores later in elementary school (e.g., first, third, and fifth grades). However, they did detect significant positive relations between teacher–child relationships and teacher-reports of students' academic achievement. Results are notable because many of the previous studies that found significant associations between teacher–child relationships and standardized student achievement failed to control for initial levels of achievement when predicting later outcomes (Burchinal, Peisner-Feinberg, Pianta, & Howes, 2002; Pianta, 1997; Pianta & Stuhlman, 2004). It is important to continue to build upon this research using a range of methods that may help address the problem posed by selection when estimating effects of teacher–child relationships on academic achievement.

It may also be important for future studies to account for systematic differences in teacher–child relationships and academic achievement that exist across schools (Kelcey, 2009; Kim & Seltzer, 2007; Singer & Willett, 1998). Ecological theories suggest that teacher–child relationship quality is likely to differ across schools (Bronfenbrenner & Morris, 1998), as are the processes by which students select into high-quality teacher–child relationships (Kim & Seltzer, 2007). For example, Hong and Raudenbush (2006) encountered such between-school variation in their study estimating the effect of kindergarten retention on achievement.

Specifically, they found students' probability of being retained was not only a function of individual characteristics but also of the student's school membership. Such a challenge highlights the importance of including school membership as a confounding covariate in cases where the school is likely to influence both the variable of interest and the outcome. For example, in the current study, it is plausible that high-achieving students in a given school are more likely to have positive teacher–child relationships due to that school's policy of allowing teachers to select children who are a better fit for certain classrooms (Hanushek & Rivkin, 2010a).

1.3. Multilevel propensity score analysis

For ethical reasons it not feasible to use the gold standard for estimating causal effects—the randomized control trial—to assign children to high and low-quality relationships with their teachers in a naturalistic school setting (Shadish, Cook, & Campbell, 2001). Thus, one popular and increasingly common method to identify comparable individuals, address selection bias, and estimate causal effects is propensity score analysis (Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007). Propensity score methods rely on a model of the treatment assignment to match individuals on the basis of similar probabilities of receiving treatment (Hill et al., 2005). Such models can be estimated using standard logistic regression, where the outcome is the treatment indicator and the predictors are all the confounding covariates. Then, for each treatment observation a match is found by choosing the control observation with the closest propensity score. Propensity score matching may be practically useful because it does not hold as strict assumptions about the linearity of the data as required for traditional regression models. Moreover, propensity score models only take into account individuals in the analysis who are included in the sample, thus limiting causal inference to the available data (Gelman & Hill, 2007; Rosenbaum & Rubin, 1983). Because this method typically uses a matching procedure, it is also helpful for identifying clear treatment and counterfactual groups about whom to interpret findings (Gelman & Hill, 2007).

Recent work has begun to use propensity score models in multilevel frameworks (e.g., Hong & Raudenbush, 2006; Kelcey, 2011). This strategy accounts for the confounding nature of school group membership by including school fixed effects in the models predicting the likelihood of experiencing the treatment (Hong & Raudenbush, 2006). By addressing school membership, we can limit bias in the estimate attributed to school differences. This method is likely appropriate in a study of teacher–child relationships and academic achievement as it is possible that the factors predictive of high-quality relationships in some schools may be relatively inconsequential in others. Under the presence of this kind of cross-level interaction on the probability of experiencing a treatment, each school likely has a different propensity equation; as a result, matching based on a uniform equation across all schools can result in misleading matches (Kim & Seltzer, 2007). By matching treatment and control students within schools, researchers can begin to address problems caused by confounding school-level variation (Hill, 2011; Hong & Raudenbush, 2006).

1.4. The current study

The current study uses a multilevel propensity score matching approach to estimate causal effects of a high-quality teacher–child relationship in kindergarten on math and reading achievement at the beginning of first grade for a sample of lower-income Black and Hispanic students ($N = 324$) attending urban elementary schools. Currently, no known studies have examined this relationship using multilevel propensity score methods. To aid comparison of these effects across extant studies, we also present results from multilevel regression models examining the effect of teacher–child relationships on academic achievement. And finally, because we chose to use a mean cut-point when operationalizing a high-quality versus lower-quality teacher–child relationship, we present a series of sensitivity analyses conducted with generalized propensity score matching procedures to examine the continuous effects of teacher–child relationships on achievement. Based on extant conceptual and empirical work, we hypothesized significant effects of high-quality teacher–child relationships on math and reading achievement.

2. Method

2.1. Participants and setting

The data for the current study are derived from the longitudinal efficacy trial of INSIGHTS into Children's Temperament (McClowry, O'Connor, Cappella, & McCormick, 2011). Twenty-two elementary schools in three inner-city school districts with students of comparable socio-demographic characteristics were partners in conducting this study. Study schools were within one standard deviation of the average elementary school size in the participating districts, and they were representative of the overall district demographic characteristics. The student populations at the schools were mainly Black (school $M = 79.13\%$) and Hispanic (school $M = 44.21\%$). The majority of the students were eligible for free or reduced price lunch (school $M = 79.97\%$). Participating schools had an average of 46.43% ($SD = .13$) of students scoring at the average level or higher on the state standardized language arts test, and 59.36% ($SD = .18$) scoring at the average level or higher on the state standardized math test.

Individual study participants included 324 children and their caregivers and 60 kindergarten teachers. The children ranged from 4 to 7 years of age at baseline ($M = 5.38$ years, $SD = .61$ years). All children were enrolled in kindergarten at baseline. Half (50%) of the children were boys. Eighty-seven percent of the children qualified for free or reduced lunch programs. Approximately 72% of children were Black, 19% were Hispanic nonBlack, and the remaining children were Biracial. A majority of the caregivers enrolling children in the study were biological mothers (84%); other caregivers included fathers (8%) and kinship guardians (7%). Caregivers ranged in age from 19 to 72 years ($M = 34.91$ $SD = 8.71$). Approximately 28% of the caregivers had

education levels less than or equivalent to a high school degree or General Education Development (GED) diploma; 26% had at least a high school degree or GED diploma; and 24% had at least some college experience.

Teacher participants included 60 kindergarten teachers (96% of whom were women). Sixty-one percent of the teachers reported their race/ethnicity as Black, nonHispanic, 10% as Hispanic/Latino nonBlack, 23% as White, and 6% as Asian or Biracial. All teachers reported having earned a bachelor's degree, and 96% of teachers had a master's degree.

2.2. Measures

The present study is particularly well-suited to a propensity score matching approach because it includes a set of baseline child- and family-level covariates that are theoretically related to both the quality of teacher–child relationships in kindergarten and children's academic achievement during the transition to first grade. Because we sought to limit threats to the internal validity of the study posed if the assumption of ignorability was not met, we decided to use all available student-level data for the current analysis. However, each confounding covariate also exhibits a theoretical and empirical association with both the treatment (the high-quality teacher–child relationship) and the outcome (academic achievement), which strengthens the predictive power of the propensity score specification (Gelman & Hill, 2007; Hill, 2011).

The dataset is also appropriate for a propensity score analysis because child information is available for three developmental time points (T1 = December/January in kindergarten, T2 = May/June in kindergarten, and T3 = October/November in first grade). In studies that seek to infer causality, characteristics that students are matched on should be assessed prior to the treatment variable (the teacher–child relationship in the present study). Then, it is important that the assessment of the treatment precede measurement of the outcome. In other words, there is a necessary temporal pattern in which the cause precedes the effect (Hill, 2011).

Although the data for the current study were drawn from a larger randomized efficacy study of a prevention program (INSIGHTS into Children's Temperament), the treatment in the current study (high-quality teacher–child relationships) is distinct from the traditional operationalization of intervention treatment (e.g., assignment to the experimental group versus the control group). Because we accounted for assignment to the INSIGHTS condition in all models we were able to identify the teacher–child relationship as the variable of interest in analyses, and we refer to it as the “treatment” when describing the methods and results for the current study (Gelman & Hill, 2007).

2.3. Teacher–child relationship quality

The 15-item teacher-reported Student–Teacher Relationship Scale (STRS; Pianta, 1992) was used to assess teacher perceptions of the quality of the teacher–child relationship at the beginning and end of the student's kindergarten year. Using a 5-point Likert scale that ranged from 1 (*definitely does not apply*) to 5 (*definitely applies*), teachers rated how applicable statements were to their current relationship with a particular child. The STRS evaluates the teacher's feelings and beliefs about the student's actions toward him or her. The items are based on attachment theory and the Attachment Q-Set (Waters & Deane, 1985). The STRS has been widely used in studies with preschool and elementary school children. It is associated with children's and teachers' classroom behaviors and correlates with observational measures of quality of the teacher–child relationship (e.g., Birch & Ladd, 1997; Howes & Hamilton, 1992; Howes & Ritchie, 1999). Additionally, STRS scores correlate with Attachment Q-Set ratings of teachers and students such that higher STRS scores are associated with more secure relationships (Howes & Ritchie, 1999). Similar to Maldonado-Carreño and Votruba-Drzal (2011) and O'Connor and McCartney (2007), we chose to work with the Total Teacher–Child Relationship Score. Possible scores ranged from 2 (lowest quality teacher–child relationship) to 10 (highest quality teacher–child relationship). The teacher–child relationship at T1 was used as a control variable in all regression models, and the T2 variable was used to operationalize the “treatment” or experience of a high-quality teacher–child relationship. In the current study, $\alpha = .94$ for the Total Teacher–Child Relationship Score at T1 and T2.

2.4. Reading and math achievement

Reading and math achievement were assessed using the raw scores from the Letter–Word Identification and Applied Problems subtests of the Woodcock–Johnson III Tests of Achievement, Form B (WJ III; Woodcock, McGrew, & Mather, 2001). The Letter–Word Identification subtest assesses letter naming and word decoding skills by asking children to identify a series of letters and words presented in isolation. The Applied Problems subtest assesses children's simple counting skills and the ability to analyze and solve mathematical word problems presented orally. The WJ III is a nationally normed and widely used achievement test with demonstrated internal consistency for children in kindergarten and first grade (NICHD Study of Early Child Care and Youth Development, 2007). Both the Letter–Word Identification and Applied Problems subtests have been shown to correlate with the Kaufman Test of Educational Achievement and the Wechsler Individual Achievement Test for elementary school-aged children (Woodcock et al., 2001). Possible scores on the Letter Word ID subtest range from 0 to 76 and possible scores on the Applied Problems subtest range from 0 to 64. Reading and math achievement measures were collected at T1 and T3. Measures from T1 were used as confounding covariates in analyses and measures from T3 were used to operationalize the outcomes.

2.5. Background measures or confounding covariates

Confounding covariates represent the variables that might influence the likelihood of having a high-quality teacher–child relationship as well as high reading and math achievement. All confounding covariates were measured at T1, which was pretreatment. Recent research suggests that only variables measured pretreatment should be used as controls in causal models. Variables measured posttreatment may be inappropriate because they may have been influenced by the treatment (Gelman & Hill, 2007; Hill et al., 2005). As described in the introduction section of this article, extant empirical research has found relations between each of the confounding covariates described below, the treatment, or the outcome.

2.5.1. Demographic confounding covariates

Parents reported on a series of demographic characteristics about themselves and their child. Child-level confounding covariates included child ethnicity (dummy coded for Hispanic or Black), child's gender (male = 1, female = 0), child age (days from birth to Time 1 assessment), and child free-lunch eligibility (eligible = 1, not eligible = 0). Parent-level confounding covariates included parent gender (parent female = 1, parent male = 0), age (in years), parent ethnicity (dummy coded for Hispanic or Black), level of education (dummy coded as 4 binary variables representing educational completion: less than high school, high school, some college, and college), whether the parent is married or not (married = 1, not married = 0), and parental work status (dummy coded as 3 binary variables representing: full-time work, part-time work, and does not work).

2.5.2. Parent involvement in elementary school

Parent involvement in children's education was assessed with the parent-reported Family Involvement Questionnaire for Elementary School (FIQ-E), an adaptation of a questionnaire originally developed for early childhood (FIQ-E; Manz, Fantuzzo, & Power, 2004). Consisting of 44 parent-reported items, the FIQ-E was developed for lower-income urban families and field tested with a large sample of African-American families. In examining the validity of the FIQ-E, Fantuzzo, Tighe, and Perry (1999) demonstrated significant correlations between the measure and documented parent volunteer experiences in school (Fantuzzo et al., 1999). The measure asks parents to report on the frequency with which they engage in a range of behaviors related to their child's schooling on a scale from 1 (*never*) to 4 (*always*). A mean score was calculated from the scale items, and possible scores thus range from 1 to 4. Cronbach's alpha in the current study was .96.

2.5.3. Child behavior problems

Behavior problems were measured with the 36-item Sutter–Eyberg Student Behavior Inventory, the teacher-report version of the Eyberg Child Behavior Inventory (Eyberg & Pincus, 1999). On a frequency scale ranging from 1 to 7 (1 = *never* to 7 = *always*), teachers reported on the frequency with which each consented child engaged in a range of problematic behaviors, such as “acts defiant when told to do something,” “has temper tantrums,” “verbally fights with other students,” and “is overactive and restless.” A mean score was calculated from the scale items, and possible scores thus range from 1 to 7. Querido and Eyberg (2003) examined the validity of the SESBI and found significant correlations between the measure and the *Conners Teacher Rating Scale—Revised* (Conners, Sitarenios, Parker, & Epstein, 1998). Cronbach's alpha in the current study was .97.

2.5.4. Child sustained attention

Data collectors assessed children's sustained attention using the Attention Sustained subtest from the Leiter International Performance Scale—Revised (Roid & Miller, 1997). Children were shown a page with pictures of a variety of objects scattered throughout and a target object at the top. They were asked to cross out as many of the objects matching the target as possible without accidentally crossing out any other objects. Children were given a limited amount of time to perform four trials (30 s for the first three trials and 60 s for the fourth) but were not scored on speed. Their performance across trials was averaged to yield two attention scores. The number of cross-outs of objects matching the target reflected the child's focused attention, while the number of cross-outs of objects not matching the target was reversed to represent the child's lack of impulsivity. Scores were standardized against a national norming sample ($M = 10, SD = 3$). The task has demonstrated high internal consistency reliability ($\alpha = .83$) for children age 5 years and good test–retest reliability ($r = .85$; Roid & Miller, 1997). The Attention-Sustained subtest has shown consistent validity, correlating highly with traditional intelligence tests (Roid & Miller, 1997).

2.5.5. Child academic competence

The raw scores on the Mathematics, Reading/Language Arts, and Critical Thinking subscales of the Academic Competency Evaluation Scale (DiPerna & Elliott, 2000) measured teacher-reported perceptions of children's academic skills and achievement-related behaviors in the winter and spring of kindergarten, and the fall of first grade. The Mathematics subscale contains 8 items, the Reading/Language Arts subscale includes 11 items, and the Critical Thinking subscale includes 9 items. For each subscale, teachers used a five-point scale (1 = *far below*, 3 = *grade level*, and 5 = *far above*) to measure students' academic skills in comparison with the grade-level expectations at their particular school. The mean score for each subscale was calculated; possible scores ranged from 1 to 5. The ACES has demonstrated validity through factor analysis and correlations with similar measures, such as the Iowa Test of Basic Skills and grade-point averages (DiPerna & Elliott, 2000). In the current study, all subscales demonstrated high levels of internal consistency (Math $\alpha = .98$, Reading/Language Arts $\alpha = .97$, and Critical Thinking $\alpha = .97$).

2.6. Procedures

2.6.1. Participant recruitment

Schools serving low-income students in three urban school districts in a large northeastern city were targeted for participation. The principal investigator and research team contacted principals in these districts to inform them about the purpose of the study and explain data collection procedures. Selection of schools occurred in three consecutive years; 23 principals agreed to participate over the three waves. One school withdrew from the study because of a principal transition, resulting in 22 schools at baseline data collection.

Kindergarten and first-grade teachers were recruited in school, small group, or individual meetings with a member of the research team. In these meetings, study goals, design, and data collection were explained. Ninety-six percent of kindergarten and first grade teachers in the 22 schools consented to participate. No kindergarten teachers withdrew from the study between the fall and spring of the first study year. All first-grade teachers completed data collection protocols at the beginning of the first-grade year.

A racially and ethnically diverse team of field staff recruited parents from the participating teachers' classrooms during fall of the first year. Parents were informed of study goals and procedures in individual meetings at the school when parents were often present (e.g., conference days and before or after school). After a parent consented, child assent was acquired through oral assenting procedures. Written materials were sent home with children, and interested parents contacted researchers for more information. Due to the intensive nature of data collection procedures, teacher burden was considered. Team members enrolled 4 to 10 children per classroom (approximately 27% of the children attending kindergarten in the targeted schools). All recruitment processes were approved by university and school district research boards.

2.6.2. Data collection

The current study uses data from three time points. Time 1 (T1) data were collected in the winter (December/January) of the kindergarten year and Time 2 (T2) data were collected in the late spring (May/June) of the kindergarten year. Time 3 (T3) data were collected in the fall (October) of the first grade year.

2.6.2.1. Parent-reports. Parents completed measures at their child's school via audio-enhanced computer-assisted self-interviewing software (Audio-CASI). This technology facilitates data collection for respondents with low literacy levels, limits socially desirable responses, and standardizes data collection procedures (Cooley et al., 1996; Couper, Singer, & Tourangeau, 2003). Parents took approximately 30 min to complete measures and received \$20 for participation.

2.6.2.2. Teacher reports. Teachers completed paper questionnaires for each consented student. The reports took teachers 1 to 2 h to complete (approximately 15 min per student). The teachers received \$50 gift cards to purchase classroom supplies each time they provided data.

2.6.2.3. Child assessments. Data collectors conducted individual child assessments with all children participating in the study. An outside consultant trained data collectors to administer the Letter-Word Identification and Applied Problems subtests of the Woodcock–Johnson III Tests of Achievement, Form B (WJ III; Woodcock et al., 2001) over a one-day training session in the fall of each year of the study (2008–2010). A graduate research assistant conducted a field reliability test with all data collectors before they were permitted to assess children and collect data.

2.6.3. INSIGHTS intervention

Following completion of baseline data collection activities in the fall of the kindergarten year, researchers used a random numbers table to assign schools to the INSIGHTS intervention or an attention-control condition. INSIGHTS is a comprehensive temperament-based intervention that integrates theory, research, and clinical applications to enhance student–teacher relationships, parenting, and children's self-regulation. The intervention provides teachers and parents with a temperament framework for supporting the individual differences of children (for information on the intervention see McClowry, Snow, & Tamis-LeMonda, 2005). Schools assigned to the attention-control condition participated in a supplemental reading program for children whose parents consented.

3. Results

3.1. Missing data analyses

For the child-level variables, there was 0% to 20% missing data across study variables. As such, we first compared students who were missing and not missing individual data points on a series of baseline characteristics, specifically, school, teacher, cohort, child ethnicity (e.g., Hispanic or Black), child's gender, child age, child free-lunch eligibility, child behavior problems, child sustained attention, child math achievement, child reading achievement, parent gender, parent age, parent ethnicity, parent education, parent marital status, and parent work status. Although we did not find substantial differences in rates of missingness between students with high and lower quality teacher–child relationships, missingness patterns between baseline variables were not completely random. Students with lower levels of parental education, parents who were not married, or had higher levels of

behavior problems, were most likely to have missing data. As such, the assumptions required for complete case analysis (or listwise deletion) were not met (Hill et al., 2005; Little & Rubin, 2002).

To achieve maximum power given the sample size ($n = 324$), a multiple imputation method (MI) was employed, and 10 separate datasets were imputed by chained equations, using STATA MICE in STATA version 12 (Little & Rubin, 2002). MI replaces missing values with predictions based on all the other information observed in the study. Unlike single imputation methods, MI accounts for uncertainty about missing data by imputing several values for each missing value, generating multiple datasets. In the current study, propensity score models and balance statistics were run 10 separate times, and final parameter estimates were generated by calculating the mean of those 10 estimates.

3.2. Bivariate correlations

We examined bivariate correlations between all study variables, excepting demographic characteristics, before proceeding with the predictive analyses to determine the extent to which confounding covariates, the treatment variable, and the outcome variables were related to one another (see Table 1).

3.3. Treatment on the treated

We were primarily interested in using multilevel models and multilevel propensity score matching techniques to estimate a “treatment on the treated” effect, or the effect of having a high-quality teacher–child relationship, comparable to those who did not experience a high-quality teacher–child relationship. We first defined a high-quality teacher–child relationship as a dichotomous variable. Teacher–child relationships that were higher than the overall mean score were coded as 1 (high-quality, treatment), and teacher–child relationships that were lower than the mean score were coded as 0 (lower quality, counterfactual). Using a binary treatment variable in this way allows for the identification of a clear counterfactual state, and an estimate that is simple to interpret, relative to the counterfactual condition (Caliendo & Kopeinig, 2008). However, because the teacher–child relationship scale is conceptually considered to be continuous, we then used a generalized propensity score method to estimate the same effect (Imai & van Dyk, 2004). Both approaches estimated the effect of a high-quality teacher–child relationship in kindergarten on achievement at the beginning of first grade, compared to what the child's achievement would have been given a lower-quality relationship with the kindergarten teacher. In all predictive analyses, an alpha of .05 was used when testing for statistical significance.

3.4. Assumptions

Multilevel propensity score matching requires a number of assumptions (Gelman & Hill, 2007; Hill, 2011). First, ignorability—a somewhat untestable assumption—must hold. In order to assume ignorability, all potentially confounding covariates must be included in the propensity score model. In the current study, including all confounding covariates is difficult, given the large number of factors theoretically and empirically linked to a teacher forming a high-quality relationship with a student. Second, there must be sufficient overlap in the distribution of propensity scores for the treatment and counterfactual groups. In other words, there must be control matches for the treatment group across the distribution of propensity scores in order to make inferences about individuals at a given propensity score. Third, the propensity score model must be appropriately specified and balance between the matched groups must be achieved. Finally, the stable unit treatment value assumption must be met. This assumption means that reading and math achievement scores for a given student cannot be dependent on the experienced teacher–child relationship of another student in the sample. An addition limitation is the school-fixed effects model assumption

Table 1
Bivariate correlations between non-demographic study variables.

Study variables	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
<i>Confounding covariates (at T1)</i>										
1. Standardized reading achievement										
2. Standardized math achievement	.48									
3. Attention sustained scaled score	.25	.36								
4. Disruptive behavior problems	-.22	.21	-.12							
5. Teacher-reported reading competence	.60	.43	.33	-.25						
6. Teacher-reported math competence	.50	.36	.27	-.28	.85					
7. Parent involvement	-.19	-.08	.03	.10	.03	-.02				
8. Teacher–child relationship	.00	.01	.05	-.16	.16	.14	.03			
<i>Treatment variable (at T2)</i>										
9. Teacher–child relationship	.02	.02	.08	.14	.16	.14	.02	.88		
<i>Outcome variables (at T3)</i>										
10. Standardized reading achievement	.13	.21	.10	.06	.10	.09	-.02	-.02	.03	
11. Standardized math achievement	.04	.07	.05	-.01	-.04	-.10	-.05	.10	.08	.18

Note. T1 = December/January in kindergarten, T2 = May/June in kindergarten, and T3 = October/November in first grade.

that any unobserved characteristics that might affect both student–teacher relationships and student math or reading achievement be time-invariant. If unobserved variables change over time in ways that are correlated with the other variables in the model, omitted variable bias still exists and ignorability is violated. We discuss the tenability of these assumptions in the Discussion section of this paper.

3.5. Multilevel regression

In the first set of analyses, unconditional two-level hierarchical linear models were run for math and reading achievement at Time 3 to determine whether there was significant between-school variation in these variables (Raudenbush & Bryk, 2002). All models were fitted in STATA 12 with XTMIXED (Rabe-Hesketh & Skrondal, 2008). XTMIXED allows one to model linear mixed-effects models (i.e., hierarchical linear models) wherein both fixed and random effects are included in the same model specification. Based on the estimates obtained from the unconditional model, intraclass correlations (ICC) were computed to represent the proportion of total variance attributed to mean differences between schools. Unconditional models suggested significant between-school variation in these data for both outcomes (Reading ICC: 9.2% and Math ICC: 13.1%). As such, a random effect was included at level 2 in all conditional models to allow the intercept to vary across schools (Raudenbush, 2009).

Conditional multilevel regression models were then run for each outcome to estimate an adjusted effect of kindergarten teacher–child relationships on first grade math and reading achievement, controlling for all confounding covariates. Regressions were run in which the treatment was operationalized as a binary variable, as well as a continuous variable. These multilevel regression models were run for all students in the sample, and no matching procedures were employed. In order to aid comparison across extant studies, we calculated effect sizes (Cohen's *d*) using procedures recommended by Feingold (2009) for calculating effect sizes in multilevel models.

Results for all models estimating treatment effects are presented in Table 3. Findings from the multilevel regression models are displayed in the first panel, followed by the results of the multilevel propensity score models in the second panel. Thus, as illustrated in the first panel of Table 3, results revealed a significant effect of teacher–child relationships on math achievement, binary treatment: $b = 1.78$, $SE = .71$, $p = .03$, $d = .34$, and continuous treatment: $b = 1.22$, $SE = .62$, $p = .02$, $d = .23$. Effects on reading achievement, however, were nonsignificant, binary treatment: $b = .12$, $SE = 1.43$, $p = .81$ and continuous treatment: $b = -.54$, $SE = 1.23$, $p = .64$. Results with both a binary treatment variable and a continuous predictor were consistent.

3.6. Multilevel propensity score matching

Because selection into a high-quality teacher–child relationship may vary across schools, and school-membership is likely to relate to outcome, failure to account for group effects in the propensity score models might lead to omitted variable bias (Kelcey, 2011; Raudenbush & Bryk, 2002; Singer & Willett, 1998). As such, two multilevel propensity score approaches were used to account for selection bias and identify clear treatment and control groups in the sample. Histograms examined prior to running models demonstrated sufficient overlap between treatment and control, meeting a key propensity score model assumption.

Propensity score models using matching with replacement were then conducted in STATA 12 using *psmatch2* (Leuven & Sianesi, 2003). The propensity score model specification included school fixed effects to allow for within-school matching. The specification for this propensity score model is as follows:

$$\text{logit}(P(Z = 1)) = \beta_0 + \alpha_i + \sum \beta_m X_{mij}$$

The propensity score was based on a vector of level-one, individual confounding covariates and the logit function was a combination of an intercept (β_0) and a series of β coefficients and individual characteristics (X_{mi}). The prediction model also included school fixed effects (α_i) in estimating students' individual propensity scores. Thus, the first step in the procedure was to estimate the probability of a child receiving treatment (i.e., a high-quality relationship with the teacher), based on a number of confounding covariates, and school membership.

Prior to estimating effects on math and reading outcomes, we first assessed the balance of the means and standard deviations of each observed covariate for the matched high versus lower quality teacher–child relationship groups. In order to improve model specification, multiple models that included different interactions and transformed covariates were tested until the groups were considered to be appropriately balanced (i.e., there were no statistically significant differences between high versus low groups across the set of observed characteristics at the $\alpha = .10$ level). Using the propensity score matching technique, a total of 112 high-quality relationship children were matched to 44 lower-quality participants. See Table 2 for a list of balance statistics, comparing the treatment group (Panel 1) with the matched control group (Panel 2) and listing the standardized mean difference and ratio of standard deviation difference between treatment and control groups for each variable (Panel 3).

The resulting weights were added to the multilevel model predicting math and reading achievement from treatment, controlling for confounding covariates, to allow for matching between the high-quality versus lower-quality groups. The composite multilevel model is expressed as:

$$Y_{ij} = \beta_0 + \tau Z_{ij} + \sum \beta_m X_{mij} + \alpha_i + \varepsilon_{ij,\text{with}} \alpha_i \sim N(0, \sigma^2_\varepsilon) \text{ and } \varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon) \text{ independent of one another}$$

Table 2

Balance of covariate means and SDs for the propensity score matched treatment and control groups.

Variable	Treatment		Matched control		T–C difference	
	M	SD	M	SD	Standardized difference	Ratio of SDs
Child boy	.44	.50	.46	.51	–0.05	0.98
Standardized reading achievement at T1	17.15	7.29	17.07	7.09	–0.07	1.08
Standardized math achievement at T1	14.03	4.53	13.97	4.64	0.08	0.95
Attention sustained scaled score at T1	9.08	8.52	9.17	8.01	–0.01	1.06
Disruptive behavior problems at T1	1.84	1.14	1.84	1.00	–0.09	1.10
Teacher-reported math competence at T1	2.77	0.62	2.77	0.62	–0.06	1.01
Teacher-reported reading competence at T1	2.74	0.75	2.77	0.69	–0.07	1.09
Parent involvement at T1	2.78	0.49	2.80	0.46	–0.05	1.08
Child Black	.74	.46	.75	.43	–0.02	1.06
Child Hispanic	.19	.40	.19	.46	0.07	0.90
Child eligible for free lunch	.90	.33	.90	.29	–0.06	1.10
Parent age	32.77	8.19	32.80	8.29	–0.01	0.99
Parent education, less than high school	.37	.48	.38	.49	–0.03	0.99
Parent education, high school diploma	.28	.45	.29	.46	–0.05	0.97
Parent education, some college	.35	.48	.31	.47	0.08	1.02
Parent education, college graduate	.15	.36	.14	.35	0.02	1.02
Parent married	.33	.48	.35	.48	–0.05	0.99
Parent Hispanic	.17	.38	.17	.38	0.02	1.02
Parent Black	.76	.45	.74	.44	0.05	1.01
Parent works full-time	.31	.33	.33	.36	–0.02	0.92
Parent works part-time	.32	.47	.30	.43	0.02	1.09

Note. T1 = December/January in kindergarten. Sample sizes: Treatment $n = 112$, Matched Control $n = 44$.

No statistically significant differences between treatment and control across the set of observed characteristics.

In this model, τ was the treatment effect, Z was the treatment assignment for student i , and $\sum \beta_m X_{mi}$ represented the individual level confounding covariates used to estimate the propensity score function. An individual error term, ε_i , was also included. School-specific constant effects of unmeasured school-level predictors were absorbed in the random effect, α_i . We assumed that the group influenced the treatment assignment of its members in a common way. In a two-level example where students are nested in schools and treatments are assigned to students under this mechanism, the school has a constant and uniform influence on each of its students.

Results from all propensity score models are displayed in the second panel of Table 3. The findings from models using matching with replacement are presented in the first two rows of Table 3. Thus, as illustrated, the results of these models identified a positive, statistically significant effect of having a high-quality teacher–child relationship in kindergarten on math achievement in first grade, $b = 3.31$, $SE = .56$, $p < .01$, $d = .63$. Presuming that the assumptions of the propensity score analyses

Table 3

Results from multilevel models estimating adjusted treatment effects of high versus lower quality teacher–child relationships in kindergarten on achievement in first grade.

Outcome variable	Multilevel model		Multilevel propensity score model	
	Treatment effect	SE	Treatment effect	SE
Matching with replacement				
Reading achievement	0.12	1.43	1.43	2.74
Math achievement	1.78	0.71 *	3.31	0.56 **
Generalized propensity score model				
Reading achievement Average Effect ^a	–0.54	1.23	1.53	1.32
Strata 1			0.63	0.98
Strata 2			3.25	1.42 *
Strata 3			1.84	1.65
Strata 4			1.83	1.73
Strata 5			0.12	1.10
Math achievement Average Effect ^a	1.22	0.62 *	2.36	0.76 **
Strata 1			1.95	0.87 *
Strata 2			3.56	1.28 **
Strata 3			3.58	1.42 **
Strata 4			1.88	0.72 *
Strata 5			0.83	0.81

Note. Sample size for propensity score models (matching with replacement), $n = 156$ students, $n = 22$ schools.

Sample size for generalized propensity score model, $n = 324$ students, $n = 22$ schools.

^a The average of the estimates from the five strata for each outcome represents the average effect in the results of the propensity score models.

** $p < .01$.

* $p < .05$.

were met, and holding constant the previously mentioned set of confounding covariates, these results suggest a positive effect of a having a high-quality teacher–child relationship in kindergarten on math achievement at the beginning of first grade, compared to students with a lower-quality teacher–child relationship.

When interpreted causally, this finding indicates that the effect of having a high-quality relationship with one's kindergarten teacher (assessed at the end of kindergarten) ranges from a 1.78 to 3.31 point higher raw score on an assessment of math achievement (WJ III Applied Problems) than would have been experienced had the relationship with the kindergarten teacher been of lower-quality. Notably, the estimate calculated in the propensity score model was larger than the estimate calculated in the multilevel model. As evident in Table 3, the estimate calculated in the propensity score model with matching with replacement was 3.31, as opposed to 1.78 in the multilevel regression model. This differential effect may be attributed to the matched control sample being advantaged, relative to the treatment group, on a number of confounding covariates (see *Caliendo & Kopeinig, 2008*). The sample mean score of the WJ III Applied Problems raw score at T3 was 19.11 ($SD = 4.49$). As such, the effect estimated in the multilevel propensity score models represents as much as a .75 standard deviation increase on the measure, relative to the counterfactual condition. However, as also evident in Table 3, these same patterns of results did not hold for the reading achievement outcome. Instead, the multilevel propensity score model with matching with replacement indicated a positive but nonsignificant relation between high-quality teacher–child relationships and reading achievement, $b = 1.43$, $SE = 2.74$, $p = .74$.

3.7. Generalized propensity score matching

Based in research suggesting that the effect of teacher–child relationships varies by level of quality (see *Crosnoe et al., 2010; O'Connor & McCartney, 2007*), a generalized propensity score matching procedure was used to calculate the same estimates when treatment was operationalized as continuous. Using *Imai and van Dyk's (2004)* framework, the effect of a continuous measure of teacher–child relationships on student achievement was estimated within each of five strata. In this approach, a regression equation is used to predict estimates for a continuous outcome from the full set of confounding covariates. Then, these estimates are divided by quintile and treatment effects are estimated separately for each of these strata. Finally, the weighted average of the five within-subclass estimates is computed to obtain an average effect. Wald tests are used to identify whether the estimates for each stratum is significantly different from 0. Although the generalized propensity score approach does use linear regression to generate estimates, by dividing predicted propensity scores up by strata and then balancing by treatment group within each group, the model is more robust to model misspecifications than a traditional regression model (*Imai & van Dyk, 2004*).

We display the findings from the generalized propensity score models in the second panel of Table 3, underneath the results for models using matching with replacement. We present both the treatment effects within each strata, and the average effect (i.e. the mean of the strata-specific effects). Results of the generalized propensity score model supported findings from the first approach, revealing a significant positive effect of a high-quality teacher–child relationship on math achievement, $b = 2.36$, $S.E. = .76$, $p < .01$, $d = .45$. Similarly, the effect estimated for the reading achievement was small and nonsignificant, $b = 1.53$, $S.E. = 1.32$, $p = .82$.

4. Discussion

Results from both multilevel propensity score matching and multilevel regression models revealed sizeable, positive impacts of high-quality teacher–child relationships in kindergarten on a standardized measure of math achievement in first grade for a low-income, racial/ethnic minority population of students attending urban schools. No effects of high-quality teacher–child relationships were detected for reading outcomes in first grade. The magnitude of the effects detected by the propensity score models was larger than those revealed by multilevel regression models. This finding may be attributed to the matched control sample being advantaged, relative to the treatment group, on a number of confounding covariates. Moreover, as reflected in the size of the standard errors, the effects of the propensity score models were more precise than those detected in the multilevel models.

4.1. Math achievement

Although this study was largely motivated by methodological concerns, there are several conceptual implications of these findings. First, the effect of the kindergarten teacher–child relationship on first grade math achievement provides evidence for the importance of teacher–child relationships in providing children with a relational foundation through which to explore new educational environments (*Belsky & Fearon, 2002*). As noted by attachment theorists, children in secure relationships are more likely to feel supported and connected to school (*Pianta, Hamre, & Allen, 2012*). Relational closeness and school connectedness, thus, foster an environment conducive to learning (*Baker, 2006; Birch & Ladd, 1998*). It may be that children who are in a secure environment with a positive teacher–child relationship are more comfortable taking the cognitive risks (e.g., possibility of failure) necessary to learn new math skills in kindergarten and at the transition to first grade (*Curby, Rimm-Kaufman, & Ponitz, 2009*).

Similarly, the findings may provide support for the influence of in-school learning for the development of numeracy and math skills. Compared with reading, math is more likely to be influenced by in-school learning, even in the earliest grades (*Grimm, 2008*). This pattern may be because parents are less likely to engage with their children in educational activities related to math and complex problem solving than they are to read with them (*Sheldon, Epstein, & Galindo, 2010*). In addition, *Crosnoe et al. (2010)* argued that because math and numeracy require complex, higher-order thinking skills, teachers who provide numeracy

instruction within supportive relationships are more likely to succeed in promoting math achievement than teachers who rely on pedagogy alone (Greenberg et al., 2003). Creating and maintaining a supportive classroom environment is especially important in under-resourced schools because more children are at-risk for academic difficulties.

The results of the current study may be especially important given an additional body of research finding that math achievement in early elementary school is a strong predictor of subsequent achievement, school completion, and college enrollment (Duncan, 2011). Duncan and Magnuson (2009) found that although persistent reading problems in kindergarten had no effect on later educational achievement and attainment, math achievement during the same time period was a significant predictor of later achievement, with effects extending into high school and college. Findings were consistent for children across income levels and urban/rural residence. These results, combined with other evidence that math achievement is critical for children's educational trajectories (see Duncan, 2011 for a review), suggest that focusing on improving teacher–child relationships matters substantially for promoting academic success.

4.2. Reading achievement

Although the absence of an effect of teacher–child relationships on reading achievement found in this study is conceptually notable, it is not novel. Indeed, a large body of research from the educational economics literature, which uses value-added modeling to estimate effects of teachers on student achievement, consistently finds small, and sometimes nonsignificant, effects of teachers on reading achievement in the early grades (see Hanushek & Rivkin, 2010b for a review). One explanation for the absence of this effect is that students' reading competencies in the early grades may be largely representative of learning that occurs outside of the school—most likely in the home (Connor, Son, Hindman, & Morrison, 2005; Foster, Lambert, Abbott-Shim, McCarty, & Franze, 2005; Storch & Whitehurst, 2001). This consideration is particularly relevant given the timing of outcome data collection in the current study: immediately following the summer break (Allington et al., 2010). Indeed, parents across racial/ethnic and economic groups report spending the majority of their home-learning time engaged in literacy activities (Ginsberg, 2012). Given recent policy and practice emphases on literacy instruction in the early grades (e.g., NCLB Reading First), it is also possible the teachers in this study spent more of their time teaching literacy and related skills than math and numeracy (Phillips, 2010). Because of this saturation of literacy activities, all students may be growing in their reading achievement regardless of the quality of their relationship with the teacher.

4.3. Methodological strengths and limitations

In addition to conceptual implications, the current study has a number of methodological strengths that help to build the research base on teacher–child relationships and achievement. Even in cases where a randomized experiment assigns children to conditions designed to promote higher-quality teacher–child relationships (e.g., Cappella et al., 2012; Pianta et al., 2012; Williford & Whittaker, 2010), there may be any number of competing “treatments” (e.g., a socio-emotional literacy curriculum and high-quality peer interactions) included in the program model. It is thus difficult to unpack the various effects to estimate a clear “treatment on the treated effect” of high-quality teacher–child relationships on academic achievement in a randomized control trial paradigm. Accordingly, the multilevel propensity score methodology is novel because it uses longitudinal data to make comparisons between groups of children who experience higher-quality teacher–child relationships and children experiencing lower quality teacher–child relationships. Because students are matched on baseline characteristics, and the quality of the teacher–child relationship is assessed after equivalence has been established, the method allows for improved causal inference.

A second methodological strength is this study's examination of selection at the school level. Because research accounting for school-level confounding has been limited, effects identified in past studies may have been overestimated and underestimated, depending on school selection processes. Models were particularly rigorous in this study, as both the treatment and outcome were included in the propensity score model specification. The treatment at Time 1 was also included in the multilevel models estimating effects on the outcome, further improving the rigor of the study design. By assessing effects in a within-group sample of urban students and schools, the study helps build the research base informing work focused on improving high-risk schools and closing achievement gaps.

However, the study does have a number of limitations. First, while the multilevel propensity score methodology represents an attempt to interpret data causally, it is impossible to identify whether the condition of ignorability has been met (Gelman & Hill, 2007). Although it is notable that child behavior problems and child academic achievement—two factors empirically related to selection into high-quality relationships (e.g., Rudasill et al., 2010)—were included in the propensity score matching, causal estimates from the propensity score models should be interpreted with caution. Moreover, if school membership does not have a uniform effect on the relation between treatment and outcome, this assumption may be further violated. Although the balance statistics for these analyses are in the range supported by the literature, it is still true that the balance between treatment and control is not perfect. As such, there may be small differences between matched treatment and control groups.

Another limitation concerns the operationalization of the treatment. We used a cut-point at the mean to separate a high-quality versus a lower-quality teacher–child relationship. Although the findings from the generalized propensity score matching approach provide some evidence that the continuous measure of the treatment variable supports the original finding, the second approach is more difficult to interpret due to the separation of the treatment and counterfactual conditions. Moreover, it would benefit future studies to identify treatment and counterfactual groups with more distinct differences when estimating effects between teacher–child relationships and achievement, perhaps by operationalizing treatment as the top 25% of relationship scores and control as the

bottom 25% of scores. Such an approach could examine possible threshold effects of teacher–child relationships on academic achievement. In the same vein, the study used the teacher–child relationship total score to operationalize the construct. However, given research showing unique relations between teacher–child Closeness and Conflict and academic achievement, future research should examine relations between these dimensions and achievement separately.

An additional limitation relates to the grouping level in the current study. Previous research with elementary-aged children suggests that classroom membership may be a more appropriate grouping mechanism than school when conducting initial matching (Raver et al., 2009). A within-classroom matching procedure was impossible for the current study due to the small sample size. However, future studies with larger samples should match children within their classrooms to address limitations to inference posed by the choice of grouping mechanism. Similarly, it is likely that overall quality of teacher instruction covaries with the teacher–child relationship. Future studies with a larger number of students nested within classrooms should include teacher-level characteristics in the propensity score specifications to tease apart effects of instruction quality and the teacher–child relationship on achievement.

Finally, although the within-group sample for the study helps to build internal validity for a subsample of particular policy interest, the results are not generalizable across a range of children and schools from diverse racial/ethnic and income backgrounds. Additional work in this area is needed to identify the extent to which these results are replicated across different populations of children, and in larger, nationally-representative datasets.

4.4. Implications for policy and practice

Bearing in mind these limitations, the current findings have implications for policy and practice. The moderate to large effect of teacher–child relationships on math achievement suggests the need to develop and test interventions for urban schools that build and maintain positive teacher–child relationships within the context of math instruction. Extant interventions often embed social–emotional and relational content into reading or literacy instruction (e.g., Brackett et al., 2009; Brown, Jones, LaRusso, & Aber, 2010); folding this content into math lessons in the early school years may be a promising approach. In addition, school psychologists may be poised to identify children with lower levels of closeness and higher levels of conflict with their teachers. Through consultation or coaching, school psychologists could support teachers in enhancing their relationships with their students. Finally, teacher education programs may benefit from educating teachers not only about academic content and pedagogical practices but also in learning strategies that build positive relationships with children in schools.

Acknowledgments

We would like to thank Marri Davis, Iyamaani Aytes, and Gise Oliviera for their project leadership, the research assistants from New York University for their assistance with data collection and management, and the school and community collaborators who contributed time and expertise to this effort. We would also like to thank Jennifer Hill, Jessica Harding, and Dana Charles McCoy for their review of early drafts of this paper.

References

- Alexander, K. L., Entwisle, D. R., Blyth, D. A., & McAdoo, H. P. (1988). Achievement in the first two years of school: Patterns and processes. *Monographs of the Society for Research in Child Development*, 53, 1–157.
- Allington, R. L., McGill-Franzen, A., Camilli, G., Williams, L., Graff, J., Zeig, J., et al. (2010). Addressing summer reading setback among economically disadvantaged elementary students. *Reading Psychology*, 31, 411–427.
- Baker, J. A. (2006). Contributions of teacher–child relationships to positive school adjustment during elementary school. *Journal of School Psychology*, 44, 211–229.
- Belsky, J., & Fearon, R. M. P. (2002). Infant–mother attachment security, contextual risk, and early development: A moderational analysis. *Development and Psychopathology*, 14, 293–310.
- Birch, S. H., & Ladd, G. W. (1997). The teacher–child relationship and children's early school adjustment. *Journal of School Psychology*, 35, 61–79.
- Birch, S. H., & Ladd, G. W. (1998). Children's interpersonal behaviors and the teacher–child relationship. *Developmental Psychology*, 34, 934–946.
- Brackett, M. A., Patti, J., Stern, R., Rivers, S. E., Elbertson, N., Chisholm, C., et al. (2009). A sustainable, skill-based approach to building emotionally literate schools. In M. Hughes, H. L. Thompson, & J. B. Terrell (Eds.), *Handbook for developing emotional and social intelligence: Best practices, case studies, and strategies* (pp. 329–358). New York, NY: Pfeiffer.
- Bronfenbrenner, U., & Morris, P. A. (1998). The ecology of developmental processes. In W. Damon, & R. M. Lerner (Eds.), *Theoretical models of human development* (5th ed.) *Handbook of child psychology*, Vol. 1. (pp. 993–1023). New York, NY: Wiley.
- Brown, L. J., Jones, S. M., LaRusso, M. D., & Aber, L. J. (2010). Improving classroom quality: Teacher influences and experimental impacts of the 4Rs program. *Journal of Educational Psychology*, 102, 153–167.
- Burchinal, M. R., Peisner-Feinberg, E., Pianta, R., & Howes, C. (2002). Development of academic skills from preschool through second grade: Family and classroom predictors of developmental trajectories. *Journal of School Psychology*, 40, 415–436.
- Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22, 31–72.
- Cappella, E., Hamre, B. K., Kim, H. Y., Henry, D. B., Frazier, S. L., Atkins, M. S., et al. (2012). Teacher consultation and coaching within mental health practice: Classroom and child effects in urban elementary schools. *Journal of Consulting and Clinical Psychology*, 80, 597–610.
- Conners, C. K., Sitarenios, G., Parker, J. D., & Epstein, J. N. (1998). Revision and restandardization of the Conners Teacher Rating Scale (CTRS-R): Factor structure, reliability, and criterion validity. *Journal of Abnormal Child Psychology*, 26, 279–291.
- Connor, C. M., Son, S. H., Hindman, A., & Morrison, F. J. (2005). Teacher qualifications, classroom practices, and family characteristics: Complex effects on first-graders' vocabulary and early reading outcomes. *Journal of School Psychology*, 43, 343–375.
- Cooley, P. C., Turner, C. F., O'Reilly, J. M., Allen, D. R., Hamill, D. N., & Paddock, R. E. (1996). Audio-CASI—hardware and software considerations in adding sound to a computer-assisted interviewing system. *Social Science Computer Review*, 14, 197–204.
- Couper, M. P., Singer, E., & Tourangeau, R. (2003). Understanding the effects of Audio-CASI on self-reports of sensitive behavior. *Public Opinion Quarterly*, 67, 385–395.
- Crosnoe, R., Morrison, F., Burchinal, M., Pianta, R., Keating, D., Friedman, S. L., et al. (2010). Instruction, teacher–student relations, and math achievement trajectories in elementary school. *Journal of Educational Psychology*, 102, 407–417.

- Curby, T. W., Rimm-Kaufman, S. E., & Ponitz, C. C. (2009). Teacher–child interactions and children's achievement trajectories across kindergarten and first grade. *Journal of Educational Psychology, 101*, 912–925.
- DiPerma, J. C., & Elliott, S. N. (2000). *Academic competence evaluation scales*. San Antonio, TX: The Psychological Corporation.
- Duncan, G. J. (2011). The importance of academic skills for preK–3rd. *A Report to the Foundation for Child Development*. Retrieved April 25, 2012 from <http://fcd-us.org/sites/default/files/Brief%20on%20Academic%20Skills%20and%20School%20Readiness%2011311.pdf>
- Duncan, G. J., & Magnuson, K. A. (2009). Can society profit from investing in early education programs? In G. J. Duncan, & K. A. Magnuson (Eds.), *Investing in Early Childhood Development* (pp. 177–188). New York, NY: Palgrave MacMillan.
- Entwisle, D. R., Alexander, K. L., & Olson, L. S. (2005). First grade and educational attainment by age 22: A new story. *The American Journal of Sociology, 110*, 1458–1502.
- Eyberg, S. M., & Pincus, D. (1999). *Eyberg Child Behavior Inventory and Sutter-Eyberg Student Behavior Inventory: Professional manual*. Odessa, FL: Psychological Assessment Resources.
- Fantuzzo, J., Tighe, E., & Perry, M. (1999). Relationships between family involvement in Head Start and children's interactive peer play. *NHSA Dialog, 3*, 60–67.
- Feingold, A. (2009). Effect sizes for growth-modeling analysis for controlled clinical trials in the same metric as for classical analysis. *Psychological Methods, 14*, 43–53.
- Foster, M. A., Lambert, R., Abbott-Shim, M., McCarty, F., & Franze, S. (2005). A model of home learning environment and social risk factors in relation to children's emergent literacy and social outcomes. *Early Childhood Research Quarterly, 20*, 13–36.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.
- Ginsberg, L. (2012). Effective strategies for teaching math to adults. In B. H. Wasik (Ed.), *Handbook of family literacy* (pp. 195–208) (2nd ed.). New York, NY: Routledge.
- Greenberg, M. T., Weissberg, R. P., O'Brien, M. U., Zins, J. E., Fredericks, L., Resnik, H., et al. (2003). Enhancing school-based prevention and youth development through coordinated social, emotional, and academic learning. *American Psychologist, 58*, 466–474.
- Grimm, K. J. (2008). Longitudinal associations between reading and mathematics achievement. *Development Neuropsychology, 33*, 410–426.
- Hamre, B., & Pianta, R. (2001). Early teacher–child relationships and the trajectory of children's school outcomes through eighth grade. *Child Development, 72*, 625–638.
- Hanushek, E. A., & Rivkin, S. G. (2010a). The quality and distribution of teachers under the No Child Left Behind Act. *Journal of Economic Perspectives, 24*, 133–150.
- Hanushek, E. A., & Rivkin, S. G. (2010b). Generalizations about using value-added measure of teacher quality. *American Economic Review, 100*, 267–271.
- Hill, J. L. (2011). Introduction to propensity score matching. *Paper presented at Topics in Policy Analysis and Evaluation meeting*. New York, NY: New York University.
- Hill, J. L., Waldfogel, J., Brooks-Gunn, J., & Han, W. (2005). Maternal employment and child development: A fresh look using newer methods. *Developmental Psychology, 41*, 833–850.
- Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Paper presented to the Department of Statistics, University of Chicago, IL*.
- Howes, C., & Hamilton, C. E. (1992). Children's relationships with caregivers: Mothers and child-care teachers. *Child Development, 63*, 859–866.
- Howes, C., Phillipsen, L. C., & Peisner-Feinberg, E. (2000). The consistency of perceived teacher–child relationships between preschool and kindergarten. *Journal of School Psychology, 38*, 113–132.
- Howes, C., & Ritchie, S. (1999). Attachment organizations in children with difficult life circumstances. *Development and Psychopathology, 11*, 251–268.
- Hughes, J. N. (2011). Longitudinal effects of teacher and student perceptions of teacher–student relationship qualities on academic adjustment. *The Elementary School Journal, 112*, 38–60.
- Hughes, J. N., Cavell, T. A., & Wilson, V. (2001). Further support for the developmental significance of the quality of the teacher–student relationship. *Journal of School Psychology, 39*, 289–301.
- Hughes, J. N., Luo, W., Kwok, O., & Loyd, L. K. (2008). Teacher–student support, effortful engagement, and achievement: A 3-year longitudinal study. *Journal of Educational Psychology, 100*, 1–14.
- Imai, K., & van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistics Association, 99*, 854–866.
- Jacob, B., & Ludwig, J. (2009). Improving educational outcomes for poor children. In M. Cancian, & S. Danziger (Eds.), *Changing poverty, changing policies* (pp. 266–300). New York, NY: Sage.
- Jerome, E. M., Hamre, B. K., & Pianta, R. C. (2009). Teacher–child relationships from kindergarten to sixth grade: Early childhood predictors of teacher-perceived conflict and closeness. *Social Development, 18*, 915–945.
- Kelcey, B. (2009). Propensity score variable selection and model type in multilevel setting sin which treatment assignment varies by school. *Paper presented at the annual meeting of the American Educational Research Association Conference, San Diego, CA*.
- Kelcey, B. (2011). Matching across schools with multilevel propensity scores. *Paper presented at the annual meeting of the American Educational Research Association Conference, New Orleans, LA*.
- Kim, J., & Seltzer, M. (2007). Causal inference in multilevel settings in which selection process vary across schools. *Working Paper, 708*, Los Angeles: Center for the Study of Evaluation (CSE).
- Ladd, G. W., Birch, S. H., & Buhs, E. S. (1999). Children's social and scholastic lives in kindergarten: Related spheres of influence? *Child Development, 70*, 1373–1400.
- Ladd, G. W., & Burgess, K. B. (1999). Charting the relationship trajectories of aggressive, withdrawn, and aggressive/withdrawn children during early grade school. *Child Development, 70*, 910–929.
- Leuven, E., & Sianesi, B. (2003). PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. *Software* (<http://ideas.repec.org/c/boc/bocode/s432001.html>).
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.) New York, NY: Wiley.
- Ly, J., Zhou, Q., Chu, K., & Chen, S. H. (2012). Teacher–child relationship quality and academic achievement of Chinese–American children in immigrant families. *Journal of School Psychology, 50*, 535–553.
- Maldonado-Carreño, C., & Votruba-Drzal, E. (2011). Teacher–child relationships and the development of academic and social skills during elementary school: A within and between child analysis. *Child Development, 82*, 601–616.
- Manz, P. H., Fantuzzo, J. W., & Power, T. J. (2004). Multidimensional assessment of family involvement among urban elementary school students. *Journal of School Psychology, 42*, 461–475.
- McClowry, S. G., O'Connor, E. E., Cappella, E., & McCormick, M. P. (2011). A preliminary examination of the efficacy of INSIGHTS in enhancing the academic learning context. *Paper presented at the annual meeting of the Society for Research on Educational Effectiveness, Washington DC*.
- McClowry, S. G., Snow, D. L., & Tamis-LeMonda, C. S. (2005). An evaluation of the effects of INSIGHTS on the behavior of inner city primary school children. *The Journal of Primary Prevention, 26*, 567–584.
- Murray, C., & Zvoch, K. (2011). Teacher–student relationships among behaviorally at-risk African American youth from low-income backgrounds: Student perceptions, teacher perceptions, and socioemotional adjustment correlates. *Journal of Emotional and Behavioral Disorders, 19*, 41–54.
- NICHD Study of Early Child Care and Youth Development (2007). Retrieved December 12, 2011, from <http://secc.rti.org>
- O'Connor, E., & McCartney, K. (2007). Examining student–teacher relationship and achievement as part of an ecological model of development. *American Educational Research Journal, 44*, 340–369.
- Phillips, D. K. (2010). On transitional space, unresolved conflicts, and an uncertain teacher education. *Teachers and Teaching: Theory and Practice, 16*, 633–644.
- Pianta, R. C. (1992). *The Student Teacher Relationship Scale*. Charlottesville, VA: University of Virginia.
- Pianta, R. C. (1997). Adult–child relationship processes and early schooling. *Early Education & Development, 8*, 11–26.
- Pianta, R. C. (1999). *Enhancing relationships between children and teachers*. Washington, DC: American Psychological Association.
- Pianta, R. C., Hamre, B. K., & Allen, J. P. (2012). Teacher–student relationships and engagement: Conceptualizing, measuring, and improving the capacity of classroom interactions. In S. L. Christenson (Ed.), *Handbook of student engagement* (pp. 365–386). New York, NY: Guildford.

- Pianta, R. C., & Stuhlman, M. W. (2004). Teacher–child relationships and children's success in the first years of school. *School Psychology Review*, 33, 444–458.
- Querido, J. G., & Eyberg, S. M. (2003). Psychometric properties of the Sutter-Eyberg Student Behavior Inventory–Revised with preschool children. *Behavior Therapy*, 34, 1–15.
- Rabe-Hesketh, S., & Skrondal (2008). *Multilevel and longitudinal modeling using Stata* (2nd ed.). College Station, TX: Stata Press.
- Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In H. Cooper, L. V. Hedges, & J. C. Valetine (Eds.), *The handbook of research synthesis* (pp. 295–315). New York, NY: Sage.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.) Thousand Oaks, CA: Sage.
- Raver, C. C., Jones, S. M., Li-Grining, C. P., Zhai, F., Metzger, M. W., & Solomon, B. (2009). Targeting children's behavior problems in preschool classrooms: A cluster-randomized controlled trial. *Journal of Consulting and Clinical Psychology*, 77, 302–316.
- Roid, G. H., & Miller, L. J. (1997). *Examiners manual: Leiter International Performance Scale–Revised*. Chicago, IL: Stoelting Co.
- Roorda, D. L., Koomen, H. M. Y., Spilt, J. L., & Oort, F. J. (2011). The influence of affective teacher–student relationships on students' school engagement and achievement: A meta-analytic approach. *Review of Educational Research*, 81, 493–529.
- Rosenbaum, P., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 4155.
- Rudasill, K. M. (2011). Child temperament, teacher–child interactions, and teacher–child relationships: A longitudinal investigation from first to third grade. *Early Childhood Research Quarterly*, 26, 147–156.
- Rudasill, K. M., Reio, T., Stipanovic, N., & Taylor, J. E. (2010). A longitudinal study of student–teacher relationship quality, difficult temperament, and risky behavior from childhood to early adolescence. *Journal of School Psychology*, 48, 389–412.
- Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. S., & Shavelson, R. J. (2007). *Estimating causal effects: Using experimental and observational designs*. Washington, DC: American Educational Research Association.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.
- Sheldon, S. B., Epstein, J. L., & Galindo, C. L. (2010). Not just numbers: Creating a partnership climate to improve math proficiency in schools. *Leadership and Policy in Schools*, 9, 27–48.
- Singer, J. D., & Willett, J. B. (1998). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.
- Spilt, J. L., Hughes, J. N., Wu, J., & Kwok, O. (2012). Dynamics of teacher–student relationships: Stability and change across elementary school and the influence on children's academic success. *Child Development*, 83, 1180–1195.
- Storch, S. A., & Whitehurst, G. J. (2001). The role of family and home in the literacy development of children from low-income backgrounds. In P. R. Britto, & J. Brooks-Gunn (Eds.), *The role of family literacy environments in promoting young children's emerging literacy skills*, Vol. 92. (pp. 53–71) San Francisco, CA: Jossey-Bass.
- Waters, E., & Deane, K. (1985). Defining and assessing individual differences in attachment relationships: Q-methodology and the organization of behavior in infancy and early childhood. In I. Bretherton, & E. Waters (Eds.), *Monographs of the Society for Research in Child Development*, 50. (pp. 41–65).
- Williford, A. P., & Whittaker, J. E. (2010). Teacher–child relationship quality: Key to improving child outcomes. Paper presented at the U.S. Department of Education and Department of Health and Human Services' early childhood meeting entitled "Early Childhood 2010: Innovation for the Next Generation," Washington, DC.
- Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Examiner's manual*. Woodcock–Johnson III Tests of Achievement. Itasca, IL: Riverside Publishing.
- Wu, J., Hughes, J. N., & Kwok, O. (2010). Teacher–student relationship quality type in elementary grades: Effects on trajectories for achievement and engagement. *Journal of School Psychology*, 48, 337–355.