Abstract Title Page

Title: Estimating causal effects of teacher-child relationships on reading and math achievement in a high-risk sample: A multi-level propensity score matching approach

Authors and Affiliations:
Meghan P. McCormick
New York University

Erin E. O'Connor
New York University

Elise Cappella
New York University

Sandee G. McClowry
New York University

Abstract

Background. A robust body of research has identified associations between positive teacher-child relationships – characterized by high levels of closeness and low levels of conflict – and children's academic achievement in elementary school (e.g. Roorda, 2012). 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 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 multi-level models to infer causal impacts of high-quality teacher-child relationships on academic achievement within this high-risk population of students and schools.

Indeed, research seeking to identify causal effects of teacher-child relationships on achievement is limited. Existing studies have typically used 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). However, 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 achievement. Complicating interpretation is the fact that associations between teacher-child relationships and achievement may actually reflect rater effects, depending on the method with which data were collected (Maldonado-Carreno & Votruba-Drzal, 2011).

Moreover, many studies have yet to account for systematic differences in teacher-child relationships and academic achievement that are likely to exist across schools (Kelcey, 2009; Kim & Seltzer, 2007; Hong & Raudenbush, 2006). Correspondingly, selection bias may originate from both student and school levels (Raudenbush, 2009; Peugh, 2010). 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.

Purpose. The current study uses a multi-level propensity score matching approach to estimate causal effects of a high-quality teacher-child relationship in kindergarten on student math and reading achievement at the transition to first grade. In this study, we compare effects from these models to multi-level regressions, and conduct sensitivity analyses of continuous effects using generalized propensity score matching procedures. We hypothesize significant effects of high-quality teacher-child relationships on math and reading achievement.

Setting. Twenty-two elementary schools in three inner-city school districts with students of comparable socio-demographic characteristics were partners in conducting this study.

Participants. Participants included n = 324 children and parents and n = 60 teachers in kindergarten. The children ranged from four to seven years of age at baseline (M = 5.38 SD = 0.61). All children were enrolled in kindergarten at baseline. Half (50%) of the children were male. Eighty-seven percent of the children qualified for free or reduced lunch programs. Approximately 72% of children were Black, 19% were Hispanic non-Black, and the remaining children were Biracial. Teacher participants included n = 60 kindergarten teachers (96% of whom were female). Sixty-one percent of the teachers reported their race/ethnicity as African American, non-Hispanic, 10% as Hispanic/Latino non-Black, 23% as White, and 6% as Asian or Biracial.

Research Design. For ethical reasons it would be impossible to use a randomized control trial to assign children to high and low-quality relationships with their teachers (Shadish, Cook,

& Campbell, 2001). One popular and increasingly common method used to identify comparable individuals and address selection bias is propensity score analysis. Propensity score methods rely on a model of the treatment assignment to identify comparable individuals on the basis of similar probabilities of receiving treatment (Hill et al., 2005). Recent work has begun to use propensity score models in multi-level frameworks (e.g. Hong & Raudenbush, 2006). This method is likely appropriate in a study of teacher-child relationships and academic achievement, as it is possible that the factors strongly predictive of high-quality teacher-child 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 and as a result, matching based on a uniform equation across all schools can result in misleading matches (Kim & Seltzer, 2007; Hill, 2012).

Data Collection. Selection of schools occurred in three consecutive years; twenty-two principals agreed to participate over the three waves. Kindergarten and first grade teachers were recruited in school, small group, or individual meetings with a member of the research team. Ninety-six percent of kindergarten and first grade teachers in the twenty-two schools consented to participate. Staff recruited parents from the participating teachers' classrooms during September and October. Child assent was acquired through oral assenting procedures. Recruitment processes were approved by university and school district research boards. Parents completed measures at their child's school via audio-enhanced computer-assisted self-interviewing software. Teachers completed paper questionnaires for each consented student

Child assessments. Data collectors conducted individual child assessments with all children participating in the study. An outside consultant trained data collectors to administer the *Woodcock-Johnson III Tests of Achievement, Form B* (Woodcock, McGrew, & Mather, 2001).

Measures. 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. In studies that seek to infer causality, characteristics that students are matched on should be assessed prior to the variable of interest (the teacher-child relationship in the present study (Hill, 2012). All variables are listed in Appendix B Table 1.

Analytic Approach. *Missing data.* There were no missing data for the school-level variables (Level 2 school indicators). However, for the child-level variables, there was 0-20% missing data across study variables. To achieve maximum power given the limited sample size (n = 324), a multiple data imputation method (MI) was employed and ten separate datasets were imputed by chained equations, using STATA MICE in STATA version 12 (Little & Rubin, 2002). Propensity score models and balance statistics were run ten separate times and final parameter estimates were generated by calculating the mean of those ten estimates.

We were primarily interested in estimating a "treatment on the treated" effect. Teacher-child relationships that were higher than the overall mean score were coded as 1 (high-quality, treatment) and relationships that were lower than the mean score were coded as 0 (lower quality, counterfactual). We then used a generalized propensity score method to estimate the same effect, assessing the treatment as a continuous measure (Imai & VanDyk, 2004).

Multi-level regression. Conditional multi-level regression models were first run for each outcome to estimate an adjusted effect of kindergarten teacher-child relationships, controlling for all confounding covariates, on first grade math and reading achievement. Regressions were run in which the treatment was operationalized as a binary variable (to compare with results from the first binary propensity score analyses), as well as a continuous variable (to compare with results from the generalized propensity score matching analysis). No matching procedures were employed.

Multi-level propensity score matching. Two multi-level 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 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:

$$logit(P(Z = 1)) = \beta_0 + \alpha_i + \sum \beta_m X_{mij} \quad (2)$$

The propensity score is based on a vector of level-one, individual confounding covariates and the logit function is a combination of an intercept (β_0) and a series of β coefficients and individual characteristics (X_{mi}). The prediction model also includes school fixed effects (α_i) in estimating students' individual propensity scores.

Prior to estimating effects on math and reading outcomes, we assessed the balance of the means and standard deviations of each observed covariate for the matched high versus lower quality teacher-child relationship groups. A total of 112 high-quality relationship children were matched to 48 lower-quality participants (see Appendix B Table 1). Resulting weights were added to the multi-level 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 multi-level model is expressed as:

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

In this model, τ is the treatment effect, Z is the treatment assignment for student *i*, and $\sum \beta_m X_{mi}$ represents the individual level confounding covariates used to estimate the propensity score function. An individual error term, ε_i, is also included. School-specific constant effects of unmeasured school-level predictors are absorbed in the random effect, α_i. 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.

Generalized Propensity Score Matching. A generalized propensity score matching procedure was used to calculate the same estimates when treatment was operationalized as continuous. Using Imai & van Dyk's (2004) framework, the effect of a continuous measure of teacher-child relationships on student achievement was estimated within each of five subclasses. The weighted average of five within-subclass estimates is computed to obtain an average treatment effect. Wald tests were used to identify whether the estimates for each strata are significantly different from 0.

Findings / Results. *Multi-level regression models.* Results revealed a significant effect of teacher-child relationships on math achievement (binary treatment: $b = 1.78$, $S.E. = 0.71$, $p < 0.05$; continuous treatment: $b = 1.22$, $S.E. = 0.62$, $p < 0.05$; see Appendix B Table 2), but not on reading achievement (binary treatment: $b = 0.12$, $S.E. = 1.43$, p = n.s; continuous treatment: $b = -0.54$, $S.E. = 1.23$, p = n.s; see Appendix B Table 2). Results with both a binary treatment variable and a continuous predictor were consistent.

Multi-level Propensity Score Models. Findings from the multi-level propensity score models suggest that there is a positive, statistically significant effect of having a high-quality teacher-child relationship in kindergarten on math achievement in first grade ($b = 3.31$, $S.E. = 0.56$, $p < 0.03$; see Appendix B Table 2). The effect of having a high-quality relationship with one's kindergarten teacher (assessed at the end of kindergarten) ranges from a 2.05 to 3.43 point higher score (average across methods) on an assessment of math achievement than would have been experienced had that relationship been of lower-quality. These same patterns of results did not hold for the reading achievement outcome. Instead, the multi-level propensity score model indicated a positive but non-significant relationship between high-quality teacher-child relationships and reading achievement (b = 1.43, *S.E.* = 2.74, p = n.s; see Appendix B Table 2).

Generalized Propensity Score Models. Results of the generalized propensity score model confirm findings from the first approach, revealing a significant positive of a high-quality teacher-child relationship on math achievement (b = 2.36, *S.E.* = 0.76, $p < 0.05$; see Appendix B Table 2). Similarly, the effect estimated for the reading outcome was small, but non-significant (b = 1.53, *S.E.* = 1.32, p = n.s.; see Appendix B Table 2).

Conclusions. The results of this study 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. However, no effects of high-quality teacher-child relations were detected for reading outcomes in first grade. Although the null finding for estimating the effects of teacher-child relationships on reading achievement is notable, it is not novel. Indeed, a large body of research from the educational economics literature has found small, and sometimes non-significant, effects on reading achievement in the early grades (see Hanushek & Rivkin, 2010).

One explanation for this null finding is that students' reading competencies in the early grades may be largely representative of learning that occurs outside of the school setting, most likely in the home (Connor et al., 2005). Math achievement in contrast, is much more likely than reading to be influenced by in-school learning, even in the earliest grades (Grimm, 2008). In addition, Crosnoe and colleagues (2010) argue that because math and numeracy require complex, higher-order thinking skills, teachers that are able to provide numeracy instruction, while providing interpersonal support to children, are likely to have more success in promoting math achievement than teachers who rely on pedagogy alone (Greenberg et al., 2003). This finding may be especially important given an additional body of research showing that math achievement in early elementary school is the strongest predictor of subsequent academic achievement, high school completion, and college enrollment (Duncan, 2011).

The current study has a number of strengths. First, the study uses a multi-level quantitative methodology to address issues posed by child selection into high-quality teacher-child relationships. The outcome and treatment at Time 1 were also included in the multi-level models estimating effects on the outcome, further improving the rigor of the study design. The study also has a number of limitations. First, it is impossible to identify whether the condition of ignorability is met (Gelman & Hill, 2007). Next, previous research of elementary-aged children suggests that classroom membership may be a more appropriate grouping mechanism than school when conducting the initial matching (Raver, et al., 2009). Although the within-group sample for the study helps to build internal validity for a subsample of particular policy interest, generalizability is limited.

The improved internal validity of the multi-level propensity school model design has implications for policy and practice. First, the moderate to large effect of teacher-child relationships on math achievement suggests that interventions should target the quality of teacher-child relationships. Interventions, however, should be multi-component and embed support for teachers' development of high-quality relationships within instructional pedagogy in mathematics.

Appendices

Not included in page count.

Appendix A. References

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.

Crosnoe, R., Burchinal, M., Keating, D., Clarke-Stewart, K. A., Morrison, F, Pianta, R, & The Eunice Kennedy Shriver National Institute of Child Health and Human Development Early Child Care Research Network (2010). Instruction, teacher-student relations, and math achievement trajectories in elementary school. *Journal of Educational Psychology, 102*, 407-417.

Duncan, G.J. (2011, January). The importance of academic skills for preK-3rd. *A Report to the Foundation for Child Development.*

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press.

Greenberg, M. T., Weissberg, R. P., O'Brien, M. U., Zins, J. E., Fredericks, L., Resnik, H., & Elias, M. J. (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*(3), 410 – 426.

Hanushek, E.A., Rivkin, S.G.(2010). Generalizations about using value-added measure of teacher quality. *American Economic Review* 100(2), 267-271.

Hill, J. L. (2012, September). Introduction to Causal Inference. *Topics in Policy Analysis and Evaluation.* Lecture conducted from New York University, New York, NY.

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*(6), 833 – 850.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: a case study of causal inference for multi-level observational data. *Seminar Series Paper, Department of Statistics, University of Chicago.*

Hughes, J. N. (2011). Longitudinal effects of teacher and student perceptions of teacher-student relationship qualities on academic adjustment. *Elementary School Journal, 112,* 38-60.

Imai, K. & van Dyk, D.A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistics Association, 99*(467), 854 – 866.

Kelcey, B. (2009, March). Propensity score variable selection and model type in multilevel setting sin which treatment assignment varies by school. Paper presented at the American Educational Research Association Conference.

Kim, J. and Seltzer, M. (2007) Causal Inference in Multilevel Settings in which Selection Process Vary across Schools. Working Paper 708, Center for the Study of Evaluation (CSE): Los Angeles.

Leuven, E. and Sianesi, B. (2003), "psmatch2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing."

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

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.*

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 (1),* 41-54.

Peugh J.L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, 48(1): 85-112.

Raudenbush, S.W. (2009). Analyzing Effect Sizes: Random-Effects Models. In Cooper, H., Hedges, L.V. & Valetine, J.C. (Eds), *The Handbook of Research Synthesis.* (pp. 295 – 315). New York, NY: Russell Sage Foundation.

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*(2), 302-316.

Roorda, D. L., H. M. Y. Koomen, J. L. Spilt and F. J. Oort. (2011). The influence of affective teacher-student relationships on students' school engagement and achievement: a meta-analytic approach. *Review of Educational Research* 81: pp. 493-529.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

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

Appendix B. Tables and Figures

Table 1.

Balance of covariate means and standard deviations for the propensity score matched versus non-matched treated (high-quality student-teacher relationship) and control (lower-quality student-teacher relationship) groups

Variable	Treatment		Unmatched Control		Matched Control		STD	Ratio
	Mean	SD	Mean	SD	Mean	SD	Diff	of SDs
Child male	0.44	0.50	0.55	0.50	0.46	0.51	-0.05	0.98
Reading achievement standard score (0 - 76)	17.15	7.29	17.07	7.09	17.07	7.09	-0.08	1.08
Math achievement standard score (0 - 63)	14.03	4.53	14.70	5.22	13.96	4.64	0.08	0.95
Attention sustained score (0 - XX)	9.08	8.52	9.38	9.05	9.17	8.01	-0.01	1.06
Disruptive behavior problems (1 - 7)	1.84	1.14	2.47	1.09	1.84	1.00	-0.10	1.10
Teacher assessment of math competence (1 - 5)	2.77	0.62	2.68	0.62	2.77	0.62	-0.06	1.01
Teacher assessment of reading competence (1 - 5)	2.74	0.75	2.66	0.70	2.77	0.69	-0.07	1.09
Parent involvement (1 - 4)	2.78	0.49	2.71	0.47	2.80	0.46	-0.05	1.08
Child Black	0.74	0.46	0.73	0.44	0.75	0.43	-0.02	1.06
Child Hispanic	0.19	0.40	0.19	0.40	0.19	0.46	0.07	0.90
Child eligible for free lunch	0.90	0.33	0.87	0.36	0.90	0.29	-0.06	1.10
Parent age	32.77	8.19	26.09	8.89	32.80	8.29	-0.01	0.99
Parent education, less than high school	0.37	0.48	0.32	0.47	0.38	0.49	-0.03	0.99
Parent education, high school diploma	0.28	0.45	0.24	0.50	0.30	0.46	-0.05	0.97
Parent education, some college	0.35	0.48	0.44	0.50	0.31	0.47	0.08	1.02
Parent education, college graduate	0.15	0.36	0.18	0.39	0.14	0.35	0.02	1.02
Parent married	0.33	0.48	0.44	0.50	0.35	0.48	-0.05	0.99
Parent Hispanic	0.17	0.38	0.18	0.39	0.17	0.38	0.02	1.02
Parent Black	0.76	0.45	0.73	0.45	0.74	0.44	0.05	1.01
Parent works full-time	0.14	0.33	0.14	0.35	0.33	0.36	-0.02	0.92
Parent works part-time	0.32	0.47	0.34	0.48	0.30	0.43	0.08	1.09

NOTE: Sample sizes: Unmatched: Treatment = 112; Control = 164; Matched: Treatment = 112; Control = 48

Table 2.
Results from multi-level models estimating adjusted treatment effects of high versus lower quality student-teacher relationships in kindergarten on achievement in first grade

	Multi-level model results			Multi-level propensity score model		
Outcome variable	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 (ATE)	-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 (ATE)	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	

** $p < 0.01$, * $p < 0.05$, † $p < 0.1$

Note: Sample size for generalized propensity score model, n = 324
Sample size for propensity score models (matching with replacement), student n = 156 (112 treatment & 44 control group members) and 22 schools