THE UNIVERSITY OF CHICAGO

ESSAYS ON THE ECONOMICS OF EDUCATION

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE IRVING B. HARRIS
GRADUATE SCHOOL OF PUBLIC POLICY STUDIES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
WILLIAM DELGADO MARTINEZ

CHICAGO, ILLINOIS
DECEMBER 2019

Dedicated with gratitude and love to my parents, Edinson and Luz Nidia, and my brothers, Edinson, Roberto, and Rony.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my entire dissertation committee. To my co-chair, Seth Zimmerman, thank you for the many hours of discussion and for challenging me and keeping me encouraged. Your guidance and advice along the way have been invaluable. To my co-chair Dan Black and committee members Damon Jones, Susan E. Mayer, and Stephen Raudenbush, thank you for generously mentoring me and so freely giving your assistance, feedback, and contributions. I would also like to thank Ariel Kalil for her continued guidance and mentorship. It was a great pleasure to work in partnership with Lauren Sartain. Cynthia Cook-Conley and Maggie DeCarlo, thank you for being such a vital resource to me all these years. I am infinitely grateful to Jonathan Hoffman for his unconditional support and fun times. Finally, I offer my infinite gratitude to my family for being pillars of love throughout my journey.

For chapters 1 and 2 of this dissertation, I and the co-authors thank the staff at Chicago Public Schools, particularly the Talent Office, and the UChicago Consortium on School Research for providing access to the data and helping us better understand the policy context. The authors gratefully acknowledge funding for this research from the Spencer Foundation.

For chapter 3, the I and co-authors gratefully acknowledge the financial support from CME Group Foundation, Paul M. Angell Family Foundation, Center for Human Potential and Public Policy, Robert R. McCormick Foundation, Heising-Simons Foundation, Overdeck Family Foundation, and the Behavioral Interventions Scholars grant 90PD0303-01-00 awarded by the Office of Planning, Research and Evaluation at the Administration for Children and Families, U.S. Department of Health and Human Services. The authors would like to thank Dan Black, Susan Dynarski, Jade Jenkins, Damon Jones, Nadav Klein, and Kate Prickett for valuable comments. We are also very grateful to the staff at the Behavioral Insights and Parenting Lab for their invaluable effort: Keri Lintz, Paula Rusca, Michelle Park Michelini, and Ana Arellano Jimenez. We also thank Dave W. Koch and numerous research assistants and assessors for their work. We thank participants at the PhD Workshop at the Harris School of

Public Policy, Advances with Field Experiments 2018, APPAM 2018 and 2019 Conferences, and APPAM Regional Student Conference 2019 for their comments. We specially thank all the Head Start preschool centers that participated in this project, their principals, teachers and staff.

ABSTRACT

My dissertation is composed of three essays on the economics of education.

The first essay, *Teachers' Comparative Advantage, School Segregation, and Educational Mobility in Chicago Public Schools*, co-authored with Lauren Sartain, examines a basic assumption that value-added (VA) models make. Specifically, this essay tests whether teacher effects are student specific. We develop and estimate a flexible multivariate VA model in which teacher effects can vary by student type and drift over time. Defining student type by race and using data on 1.7 million observations, we employ quasi-experimental and holdout strategies that exploit teacher switching in Chicago Public Schools. We find evidence of student-specific teacher effects, thus rejecting the homogeneity assumption in VA models. Multivariate teacher effects naturally create comparative advantage—that is, teachers are more effective with specific student types than other types. We characterize teachers' comparative advantage and relate it with educational mobility and student-teacher racial match effects. We also discuss the implications of deselecting teachers based on univariate rather than multivariate VA when schools are segregated.

The second essay, *Evaluator Bias: Unpacking the Relationship between Classroom Characteristics and Classroom Observation Ratings*, co-authored with Lauren Sartain and Andrew Zou, examines the association between within-teacher changes in classroom composition and teacher observation ratings. We find that having higher-achieving students, less disruptive students, and fewer disadvantaged students, even when the teacher remains in the same school, is associated with better observations. We, then, simulate the distribution of teacher performance under a hypothetical policy that adjusts observation ratings for classroom characteristics, more akin to value-added measures. 30 percent (8.6 percent) of teachers who would otherwise be in the bottom 5th (10th) percentile of the unadjusted rating distribution are above that percentile on the adjusted distribution.

The last essay, *The MPACT Initiative: Using Behavioral Tools to Increase Children's Early Math Skills*, co-authored with Susan E. Mayer and Ariel Kalil, tests the effectiveness

of a low-cost treatment on parental engagement in math activities and a child's early math skills. This chapter specifically asks (i) whether providing parents with information and materials in the form of a math activity booklet increases children's math skills, and (ii) whether a behaviorally informed treatment designed to overcome present bias and increase parents' use of the materials affects children's math skills beyond providing the information and materials alone. MPACT is a 12-week randomized control trial with more than 1,400 parents of preschool-age children attending 29 Head Start Centers in the City of Chicago. We collected data from parent surveys, teacher surveys, time-preference tasks, and children's math test scores at baseline and immediately following the 12-week intervention. Child assessments in addition are collected at 6 and 12 months postinte rvention. The results in this chapter are for short-term treatment impacts (12-week). Results show an increase in children's engagement with math activities at home but no positive treatment impacts on test scores at 12 weeks. Heterogeneity analysis suggests that Spanish-speaking families benefit more from the intervention than other parents. We discuss possible explanations for the absence of positive treatment effects, including spillover effects.

CHAPTER 1

TEACHERS' COMPARATIVE ADVANTAGE, SCHOOL SEGREGATION, AND EDUCATIONAL MOBILITY IN CHICAGO PUBLIC SCHOOLS

Co-author: Lauren Sartain

1.1 Introduction

School districts are employing value-added (VA) models to estimate the average contribution of teachers on student learning and incorporating these estimates in high-stake decisions regarding teachers' retention, dismissal, and compensation. Studies have shown that students assigned to high-VA teachers are more likely to graduate, go to college, and earn more as adults (Chetty et al., 2014b; Rothstein, 2017).

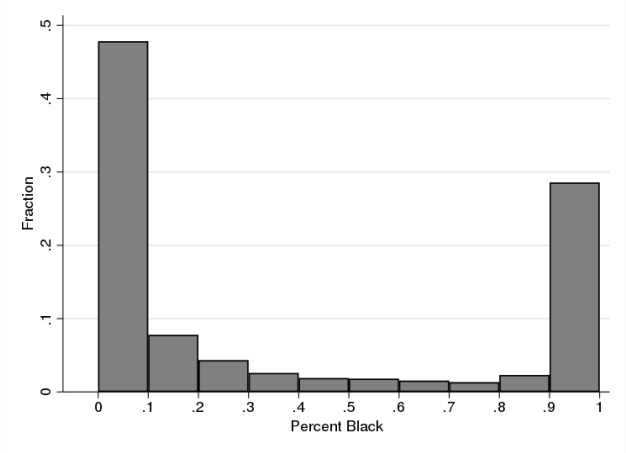
A limitation of VA models is that they assume homogeneous or univariate teacher effects and thus ignore teachers' comparative advantage, the ability to increase the outcomes of certain student populations at a lower relative opportunity cost. The concept of comparative advantage originated in the international trade literature to explain the benefits of trade and specialization. Even countries that produce goods at a lower cost than other countries would gain from trading by specializing in the production of goods with lower relative marginal cost. In educational settings, it could be possible that a more efficient outcome—that is, higher student learning—could be achieved if teachers were matched to classrooms based on their comparative advantage, rather than on their average VA.

This chapter examines the homogeneity assumption and tests whether teacher effects are student specific. Specifically, our first research question is (i) whether teacher effects vary by student's race. Student-specific effects or multivariate effects, as we call it in this chapter, naturally create comparative advantage because some teachers would have a larger impact on one student type relative to another. Our second research question is (ii) to what

extent teachers' comparative advantage relate to students' educational mobility and student-teacher racial match. We are particularly interested in race because of the racial disparities in education widely documented in the literature (Fryer and Levitt, 2006, 2013) and because of the recent trends of school resegregation (Clotfelter et al., 2006b; Lutz, 2011). To answer our research questions, we use administrative data from Chicago Public Schools (CPS), the third largest school district in the U.S., that has information on students' characteristics and test scores spanning from 2008–09 to 2016–17 school years. We use math and reading test scores as our main student outcomes.

We face two main challenges to study multivariate teacher effects by student's race. First, schools in CPS are highly segregated, so many teachers teach only one type of student. One could not be able to estimate the VA on black students for teachers who only taught to non-black students and vice versa. See, for example, Figure 1.1, which shows the distribution of the percentage of black students in CPS classrooms. The plot has a "U" shape, indicating that many classrooms have either very few, or a large proportion of black students. We find a non-trivial amount of teachers with a single type of student.

Figure 1.1: Distribution of the Percentage of Black Students in Classrooms



Notes: This figure shows the distribution of the percentage of black students in CPS classroom. Each bin is 10 percentage-points wide.

The second challenge is that multivariate teacher VA, or VA by student's race, has not been validated. That is, we do not know whether estimates of VA on blacks and non-blacks

actually reflect the causal effect of teachers on those student subgroups. Chetty et al. (2014a); Bacher-Hicks et al. (2014); Rothstein (2017); Kane and Staiger (2008); Kane et al. (2013) provide evidence that the *average* VA predicts changes in student test scores; however, their conclusions may not hold for multivariate effects. It could be possible that teachers have a univariate effect on students, irrespective of their characteristics. It could also be possible that estimates of multivariate effects are noisy (due to sample splitting into student groups) and thus their relationship with student test scores would be null.

To overcome these challenges, we first propose a flexible VA model that allows us estimate multivariate teacher effects, even for teachers who never taught some student types. We build on Chetty et al. (2014a) and relax the homogeneity assumption. We further allow teacher effects to be correlated with each other—that is, VA on black students can provide information on VA on non-black students. We show that our multivariate VA model is more general than previous models proposed in the literature and that it can also be applied when teachers affect multiple student outcomes (e.g., test scores and behaviors).

Next, we test for the existence of multivariate effects by investigating whether student test scores are only affected by own-type VA rather than other-type VA. Put in different words, black students' test scores should be affected by their teachers' VA on black students and not by his or her VA on non-black students. We exploit teacher turnover and provide quasi-experimental evidence that type-specific teacher effects matter. We provide additional evidence using holdout methods that predict the effects of teachers who switch across different classroom types. If homogeneity is assumed when in reality effects are multivariate, traditional VA models would make inaccurate predictions for this type of switchers. On the other hand, a multivariate VA model would make more accurate predictions. Our findings indicate that teacher effectiveness varies by student's race and that only own-type VA explains test scores. Our findings of multivariate teacher effects make the concept of comparative advantage policy relevant.

With our estimates of multivariate teacher VA, we construct teachers' comparative ad-

vantage as the difference between VAs (e.g., VA on blacks minus VA on non-blacks). We characterize comparative advantage and relate it with educational mobility. Educational mobility is defined as the change in ranking at the start and end of schooling, and we find that students who move upward the test score distribution are assigned to teachers with comparative advantage on their type while progressing through school.

Additionally, we conduct a policy simulation that lays off teachers based on their *mean* VA, mirroring what school districts do and what previous studies have suggested. We find that using the *mean* VA rather than the multivariate VA may mistakenly classify teachers as low performers when they actually are not. Thirty-three percent (31 percent) of teachers who are in the bottom 5th (10th) percentile would otherwise be above that percentile under heterogeneous effects (i.e., either their VA on black or non-black students is above that percentile).

This chapter is related to various strands within the economics of education literature. It relates to studies that validate VA estimates (Kane and Staiger, 2008; Kane et al., 2013; Chetty et al., 2014a; Rothstein, 2017; Bacher-Hicks et al., 2014), studies that estimate teacher effects by student's characteristics (Aaronson et al., 2007; Loeb et al., 2014; Lockwood and McCaffrey, 2009; Konstantopoulos, 2009; Milner IV, 2006; Dee, 2004; Aucejo, 2011), research on student-teacher racial match (Gershenson, 2016; Egalite and Kisida, 2018; Egalite et al., 2015), and research on match quality (Jackson, 2018).

This chapter makes four major contributions to the literature. The first contribution is methodological by extending VA models to allow for multivariate teacher effects. Most papers assume homogeneous and fixed teacher effects, but to the best of our knowledge, we are the first to propose and estimate a VA model with multivariate teacher effects that can vary over time. Second, we characterize the marginal and joint distributions of teacher VA on black and non-black students and document the extent of comparative advantage in an educational setting. Third, to the best of our knowledge, we are the first to provide quasi-experimental evidence on multivariate teacher effects using more than 1.7 million observations. Most

papers are observational, and a few are experimental with small samples. Fourth, we employ holdout methods to validate our VA estimates and compare them with estimates resulting from homogeneous VA models.

This chapter is organized as follows. Section 1.2 develops our multivariate VA model and estimators. Section 3.2.4 describes the CPS data and presents summary statistics. In Section 1.4, we present our VA estimates and characterize their probability distributions. Section 1.5 describes the teacher turnover quasi-experiment to test for multivariate effects and presents results. In Section 1.6, we explain the holdout sample approach to validate VA estimates. The next section presents some robustness checks. We then estimate teachers' comparative advantage and show its relationship with educational mobility. Section 1.9 briefly discusses policy implications and next steps. We conclude in the last section.

1.2 Analytical Framework

This section develops our multivariate VA model. Our methodology differs from previous VA models in two ways. First, we follow Chetty et al. (2014a) and allow teacher effects to drift over time. Most VA models assume constant teacher effects that do not vary over time. Second, we relax the homogeneity assumption. We allow teacher effects to vary by student types and let these effects be correlated with each other. We will see that previous VA models are special cases of our multivariate model.

1.2.1 Model

Let i index students, j teachers, and t years. Let $k(i)$ denote student i's type, which we denote as k because a student can only belong to a single type. The outcome of student i, whose type is k, taught by teacher j in year t is given by:

$$A^*_{it} = X'_{ikt}\beta_k + \nu_{ijkt} \tag{1.1}$$

where

$$\nu_{ijkt} = \mu_{jkt} + \theta_{jkt} + \varepsilon_{ijkt} \tag{1.2}$$

X_{ikt} is a $K+1$ vector of student characteristics, including demographic variables and baseline outcome ($A^*_{ik,t-1}$). ν_{ijkt} is the error term composed of teacher effect on student type k, μ_{jkt}, type-specific classroom shock, θ_{jkt}, and idiosyncratic student-level error, ε_{ijkt}. Teacher effects μ_{jkt} may fluctuate over time. Type-specific classroom shocks θ_{jkt} are idiosyncratic shocks that equally affect all students of type k. An example of this type of shock could be stereotype threat if it arises in certain years or news of police brutality against a black teenager. Note that the coefficient β_k varies by student type.

Assumption 1 (Joint stationarity of teacher effects) *Each teacher effect by student type is stationary and their cross-covariances depend only on the time difference:*

$$\mathbb{E}[\mu_{jkt}|k,t] = 0 \tag{1.3}$$

$$Var(\mu_{jkt}) = \sigma^2_{\mu_k} \tag{1.4}$$

$$Cov(\mu_{jkt}, \mu_{jk,t+s}) = \sigma_{\mu_k s} \tag{1.5}$$

$$Cov(\mu_{jkt}, \mu_{jmt}) = \sigma_{\mu_k \mu_m} \tag{1.6}$$

$$Cov(\mu_{jkt}, \mu_{jm,t+s}) = \sigma_{\mu_k \mu_m s} \tag{1.7}$$

$$Cov(\mu_{jk,t+s}, \mu_{jmt}) = \sigma_{\mu_k s \mu_m} \tag{1.8}$$

for all $m \neq k$ and for all t, $s > 0$.

Assumption 1 implies that the mean and variance of teacher effects on a student type does not vary in time. The means are centered at 0, which is a normalization. They could also be interpreted as the priors in a Bayesian sense. The variances are allowed to differ by student type. The teacher effects can be correlated with itself across time and with teacher VA on other types. The auto-covariance of teacher VA across any pair of years depends only on the amount of time between those years. The cross-covariance between teacher VA on

different types also depends only on the amount of time between those years. Note that the cross-covariance of teacher effects is not interchangeable when one variable is leading rather than the other—that is, $Cov(\mu_{jkt}, \mu_{jm,t+s}) \neq Cov(\mu_{jk,t+s}, \mu_{jmt})$.

A concern about Assumption 1 is that the variance-covariance structure of teacher effects may be different for novice and experienced teachers. Studies find that teachers' effectiveness rapidly grow early in their careers, but gains for each additional year of experience are modest later in their careers (Rockoff, 2004; Papay and Kraft, 2015). As a robustness check, we restrict the sample to experienced teachers whose effects are relatively stable over time.

Assumption 2 (Stationarity of unobserved error term) *Let $\tilde{\varepsilon}_{ijkt} = \theta_{jkt} + \varepsilon_{ijkt}$ be the unobserved error term that is unrelated to teacher effects. This error term is idiosyncratic and follows a stationary process:*

$$\mathbb{E}[\tilde{\varepsilon}_{ijkt}|k,t] = 0 \tag{1.9}$$

$$Var(\tilde{\varepsilon}_{ijkt}) = \sigma^2_{\varepsilon} + \sigma^2_{\theta_k} \tag{1.10}$$

$$Cov(\tilde{\varepsilon}_{ijkt}, \tilde{\varepsilon}_{ijmt}) = \sigma_{\theta_k\theta_m} \tag{1.11}$$

for all $m \neq k$ and for all t.

Assumption 2 requires that mean and variance of the unobserved error term do not vary across years. Additionally, the within-year cross-correlation is constant as well. In other words, individual-level shocks are i.i.d. while specific classroom shocks are correlated within year.

Given this model, the parameters to estimate are the regression coefficients β_k for all k; teacher effects μ_{jk} for all j and k; variances $\sigma^2_{\mu_k}$, σ^2_{ε}, $\sigma^2_{\theta_k}$ for all k; auto-covariance $\sigma_{\mu_k s}$ for all k and $s > 0$; and cross-covariances $\sigma_{\mu_k\mu_m 0}$, $\sigma_{\mu_k\mu_m s}$, $\sigma_{\mu_k s\mu_m}$, and $\sigma_{\theta_k\theta_m}$ for all $m \neq k$ and $s > 0$.

Note that the model above can be modified to include the case with multiple student outcomes, as in Jackson (2018); Petek and Pope (2016); Kraft (2019); Blazar and Kraft (2017);

Gershenson (2016); Koedel (2008); Ladd and Sorensen (2017). The dependent variable A^*_{ikt} would indicate the outcome k of student i taught by teacher j in year t. For instance, A^*_{i1t} may refer to test scores and A^*_{i2t} to behavioral outcomes. Under this case, Assumption 2 should be modified to allow correlated errors within each student. In particular, Eq. 1.11 would become $Cov(\tilde{\varepsilon}_{ijkt}, \tilde{\varepsilon}_{ijmt}) = \sigma_{\varepsilon_k \varepsilon_m} + \sigma_{\theta_k \theta_m}$. The model can be generalized further to allow both multiple outcomes and multiple student types. The identification strategy is similar in either of these cases.

1.2.2 Identification of teacher VA

Our approach to estimate teacher VA closely follows Chetty et al. (2014a), except that teacher effects vary across student types and these effects are correlated. The estimator of teacher VA in year t on student type k is based on mean test scores of type-k students in prior classes and mean test scores of other student types in prior classes. To illustrate the identification strategy, we assume that each teacher teaches one classroom per year, teachers have students of all types in their classrooms, the composition of students is the same across classrooms and years, and the number of years with available data for all teachers are the same. In our empirical estimation, we account for differences in class size, classroom compositions, and number of years a teacher appears in the data.

Our estimator is constructed in several steps. First, we estimate the coefficient β_k for each k student subgroup to construct test score residuals adjusting for observables. We divide the sample by student type, and for each subsample we regress test scores A^*_{ikt} on X_{ikt} using within-teacher variation. The OLS regression is

$$A^*_{ikt} = X'_{ikt}\beta_k + \alpha_j + \epsilon_{ijkt} \tag{1.12}$$

where α_j is a teacher fixed effect.

Second, we construct test scores residual as

$$A_{ikt} \equiv A^*_{ikt} - X'_{ikt}\hat{\beta}_k. \tag{1.13}$$

This residual is composed of teacher effect, classroom-level error, and individual-level error: $A_{ikt} = \mu_{jkt} + \theta_{jkt} + \varepsilon_{ijkt}$.

Third, we compute the average of the residuals at the teacher-level for each student subgroup. The mean residual test score of teacher j's students of type k is

$$\bar{A}_{jkt} = \frac{1}{n_k} \sum_{i \in \{i: j(i,t)=j \ \& \ k(i)=k\}} A_{ikt},$$

where n_k is the number of students in the classroom who belong to group k.[1]

Fourth, we estimate the best linear predictor of mean test score residuals in year t for type k based on mean test score residuals of the same type in prior years and mean test score residuals of the other student types in prior years. Let $\boldsymbol{A}^{-t}_{jk} = \left(\bar{A}_{jk1}, \ldots, \bar{A}_{jk,t-1}\right)'$ denote the $(t-1) \times 1$ vector of mean residuals in teacher j's classes prior to year t for students in subgroup k. Let $\boldsymbol{A}^{-t}_{j} = (\boldsymbol{A}'^{-t}_{j1}, \ldots, \boldsymbol{A}'^{-t}_{jK})'$ denote the $K(t-1) \times 1$ vector that stacks all mean test score residuals prior to year t for all student subgroups. On the same hand, let $\boldsymbol{A}_{jt} = (\bar{A}_{j1t}, \ldots, \bar{A}_{jKt})'$ denote the $K \times 1$ vector that contains mean residual scores across student types in teacher j's classrooms in year t. Our estimator of teacher j's effects on student types is the best linear predictor of $\boldsymbol{A}_{jt}$ based on prior scores $\boldsymbol{A}^{-t}_{j}$. It is given by:

$$\hat{\boldsymbol{\mu}}_{jt} \equiv \mathbb{E}[\boldsymbol{A}_{jt} | \boldsymbol{A}^{-t}_{j}] = \boldsymbol{\psi}' \boldsymbol{A}^{-t}_{j} \tag{1.14}$$

The $K(t-1) \times K$ matrix $\boldsymbol{\psi}$ contains reliability weights. These weights come from

1. If teachers teach multiple classrooms, one would collapse the data to the classroom level and then construct teacher-level averages as the precision-weighted average of their classroom-level residuals. The precision weights for classroom c would be $h_{ckt} = \frac{1}{\sigma^2_{\theta_k} + \frac{\sigma^2_\varepsilon}{n_{ckt}}}$, which is the inverse of the variance of the estimate of a teacher's effect from class c on student type k.

minimizing the mean-squared error of the forecasts of test scores:

$$\boldsymbol{\psi} = \arg\min \sum_{j} \left(\boldsymbol{A}_{jt} - \boldsymbol{\psi}' \boldsymbol{A}_{j}^{-t} \right)' \left(\boldsymbol{A}_{jt} - \boldsymbol{\psi}' \boldsymbol{A}_{j}^{-t} \right). \tag{1.15}$$

This minimization problem has a system of K equations, where the dependent variables are $\bar{A}_{j1t}, \ldots, \bar{A}_{jKt}$ and the explanatory variables are $\bar{A}_{j1,1}, \ldots, \bar{A}_{jK,t-1}$, which are the same variables across equations. The resulting coefficients $\boldsymbol{\psi}$ are equivalent to those obtained from seemingly unrelated regressions (SUR). They are also equivalent to the coefficients from a vector autoregression (VAR) with $t-1$ lags. Additionally, the estimated coefficients are equivalent to those obtained from separate OLS regressions because the explanatory variables are identical across equations (Zellner, 1962).

The resulting reliability weights are

$$\boldsymbol{\psi} = \boldsymbol{\Gamma}^{-1} \boldsymbol{\gamma} \tag{1.16}$$

where $\boldsymbol{\Gamma}$ is the $K(t-1) \times K(t-1)$ variance-covariance matrix of $\boldsymbol{A}_{j}^{-t}$ and $\boldsymbol{\gamma}$ is the $K(t-1) \times K$ covariance matrix between $\boldsymbol{A}_{j}^{-t}$ and $\boldsymbol{A}_{jt}$. In particular, the k-th column of $\boldsymbol{\psi}$, which contains the reliability weights for teacher VA on type k, is

$$\boldsymbol{\psi}_k = \boldsymbol{\Gamma}^{-1} \boldsymbol{\gamma}_k \tag{1.17}$$

where

$$\boldsymbol{\gamma}_k = \begin{pmatrix} \boldsymbol{\phi}_{k,1} \\ \vdots \\ \boldsymbol{\phi}_{k,K} \end{pmatrix}$$

with the m-th element being $\boldsymbol{\phi}_{k,m} = (Cov(\bar{A}_{jkt}, \bar{A}_{jm1}), \ldots, Cov(\bar{A}_{jkt}, \bar{A}_{jm,t-1}))'$, and

$$\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\Gamma}_{11} & \cdots & \boldsymbol{\Gamma}_{1K} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Gamma}_{K1} & \cdots & \boldsymbol{\Gamma}_{KK} \end{pmatrix}.$$

The m-th diagonal block of $\boldsymbol{\Gamma}$, $\boldsymbol{\Gamma}_{mm}$, is a $(t-1) \times (t-1)$ symmetric matrix equal to

$$\boldsymbol{\Gamma}_{mm} = \begin{pmatrix} Var(\bar{A}_{jm1}) & \cdots & Cov(\bar{A}_{jm1}, \bar{A}_{jm,t-1}) \\ \vdots & \ddots & \vdots \\ Cov(\bar{A}_{jm,t-1}, \bar{A}_{jm1}) & \cdots & Var(\bar{A}_{jm,t-1}) \end{pmatrix},$$

the mn-th off-diagonal block of $\boldsymbol{\Gamma}$, $\boldsymbol{\Gamma}_{mn}$, is a $(t-1) \times (t-1)$ nonsymmetric matrix equal to

$$\boldsymbol{\Gamma}_{mn} = \begin{pmatrix} Cov(\bar{A}_{jm1}, \bar{A}_{jn1}) & \cdots & Cov(\bar{A}_{jm1}, \bar{A}_{jn,t-1}) \\ \vdots & \ddots & \vdots \\ Cov(\bar{A}_{jm,t-1}, \bar{A}_{jn1}) & \cdots & Cov(\bar{A}_{jm,t-1}, \bar{A}_{jn,t-1}) \end{pmatrix},$$

and $\boldsymbol{\Gamma}_{mn} = \boldsymbol{\Gamma}'_{nm}$. The elements of $\boldsymbol{\gamma}_k$ and $\boldsymbol{\Gamma}$ are organized in blocks because of the stacked nature of the vector $\boldsymbol{A}_j^{-t}$. Appendix Section A.1 shows how the components of the reliability weight matrixes can be expressed in terms of variances, auto-covariances, and cross-covariances of teacher effects, individual shocks, and classroom shocks under Assumptions 1 and 2.

The fifth and last step of our multivariate estimators is to use the estimates of reliability weights $\boldsymbol{\psi}$ to predict teacher effects on student subgroups in period t, $\hat{\boldsymbol{\mu}}_{jt}$, as in Eq. 1.14.[2] The prediction is a leave-one-out forecast because the year t is omitted. It is also a shrinkage estimator because the reliability weights are lower than one, thus shrinking estimates toward

2. It is worth highlighting that reliability weights vary depending on the number of years of available information. Take, for example, a teacher with one single year of past data and another teacher with two years of prior information. Not only the weights for the latter teacher include both the prior year and the year before, but also the weights on the prior year differ between the two teachers.

the global mean of zero. It can also be interpreted as an Empirical Bayes estimator if teacher effects follow a multivariate normal distribution and individual and classroom shocks follow independent normal distributions.

1.2.3 Special cases

Previous VA models are special cases of our multivariate VA model. We consider three cases: (i) teacher effects and classroom shocks are uncorrelated across types, (ii) teacher effects are fixed over time but can be correlated across types, and (iii) both scenarios (i) and (ii) together.

Scenario (i): Teacher effects and classroom shocks are uncorrelated across types

In the first scenario, uncorrelated teacher effects and classroom shocks imply that the cross-covariances are equal to 0—that is, $Cov(\mu_{jkt}, \mu_{jms}) = 0$ and $Cov(\theta_{jkt}, \theta_{jms}) = 0$ for all $m \neq k$ and for all t, s. The variance-covariance matrix, $\mathbf{\Gamma}$, becomes a block diagonal matrix and the vector $\boldsymbol{\gamma}_k$ is zero everywhere except for $\phi_{k,k}$. The reliability weights for teacher j's VA on student type k in year k (Eq. 1.17) becomes

$$\boldsymbol{\psi}_k = \mathbf{\Gamma}_{kk}^{-1}\phi_{k,k}.$$

As a result, the best linear predictor is

$$\hat{\mu}_{jkt} = \boldsymbol{\psi}_k' \boldsymbol{A}_{jk}^{-t}.$$

Only the prior values of mean score residuals for k ($\boldsymbol{A}_{jk}^{-t}$) matter to predict teacher VA on this subgroup. This is equivalent to separately estimating teacher effects with different subsamples.

If $K = 1$, we return to the case of homogeneous or univariate teacher effects since students

are treated as belonging to a single group. This is the case considered in Chetty et al. (2014a).

Scenario (ii): Teacher effects are fixed but can be correlated across types

In the second scenario, where teacher effects are fixed but can be correlated, the prediction of teacher j's effect in on students of type k in year t is

$$\hat{\mu}_{jkt} = \boldsymbol{\psi}_k' \bar{\boldsymbol{A}}_j^{-t}$$

where $\bar{\boldsymbol{A}}_j^{-t} = (\bar{A}_{j1}^{-t}, \ldots, \bar{A}_{jK}^{-t})'$ and $\bar{A}_{jk}^{-t} = \frac{1}{t-1}\sum_{s=1}^{t-1} \bar{A}_{jks}$ is the mean residual test score of type k in classes taught by teacher j in years prior to t. The reliability weight becomes

$$\boldsymbol{\psi}_k = \tilde{\boldsymbol{\Gamma}}^{-1}\tilde{\boldsymbol{\gamma}}_k$$

where $\tilde{\boldsymbol{\gamma}}_k = (Cov(A_{jkt}, \bar{A}_{j1}^{-t}), \cdots, Cov(A_{jkt}, \bar{A}_{jK}^{-t}))'$ and

$$\tilde{\boldsymbol{\Gamma}} = \begin{pmatrix} Cov(\bar{A}_{j1}^{-t}, \bar{A}_{j1}^{-t}) & \cdots & Cov(\bar{A}_{j1}^{-t}, \bar{A}_{jK}^{-t}) \\ \vdots & \ddots & \vdots \\ Cov(\bar{A}_{jK}^{-t}, \bar{A}_{j1}^{-t}) & \cdots & Cov(\bar{A}_{jK}^{-t}, \bar{A}_{jK}^{-t}) \end{pmatrix}.$$

Two specific cases are embedded in the second scenario. The first specific case assumes $K = 2$. Under the stationarity assumptions, the reliability weight for the first teacher effect simplifies to

$$\begin{aligned}\boldsymbol{\psi}_1 =& \frac{\left[\sigma_{\mu_2}^2 + \left(\sigma_{\theta_2}^2 + \sigma_\epsilon^2/n_2\right)/t-1\right]\sigma_{\mu_1}^2 - \left(\sigma_{\mu_1\mu_2} + \sigma_{\theta_1\theta_2}/t-1\right)\sigma_{\mu_1\mu_2}}{\left[\sigma_{\mu_1}^2 + \left(\sigma_{\theta_1}^2 + \sigma_\epsilon^2/n_1\right)/t-1\right]\left[\sigma_{\mu_2}^2 + \left(\sigma_{\theta_2}^2 + \sigma_\epsilon^2/n_2\right)/t-1\right] - \left(\sigma_{\mu_1\mu_2} + \sigma_{\theta_1\theta_2}/t-1\right)^2} + \\ & \frac{\left[\sigma_{\mu_1}^2 + \left(\sigma_{\theta_1}^2 + \sigma_\epsilon^2/n_1\right)/t-1\right]\sigma_{\mu_1\mu_2} - \left(\sigma_{\mu_1\mu_2} + \sigma_{\theta_1\theta_2}/t-1\right)\sigma_{\mu_1}^2}{\left[\sigma_{\mu_1}^2 + \left(\sigma_{\theta_1}^2 + \sigma_\epsilon^2/n_1\right)/t-1\right]\left[\sigma_{\mu_2}^2 + \left(\sigma_{\theta_2}^2 + \sigma_\epsilon^2/n_2\right)/t-1\right] - \left(\sigma_{\mu_1\mu_2} + \sigma_{\theta_1\theta_2}/t-1\right)^2}.\end{aligned} \tag{1.18}$$

This formula coincides with the weights derived by Lefgren and Sims (2012) in their equation (7). Our equation 1.18 is their case with i.i.d individual shocks of their model.

The second specific case assumes i.i.d. of classroom shocks (in addition to i.i.d. of individual shocks). Under this and the stationarity assumptions, the cross- and auto-covariances are simplified to $Cov(A_{jkt}, \bar{A}_{jm}^{-t}) = Cov(\bar{A}_{jk}^{-t}, \bar{A}_{jm}^{-t}) = \sigma_{\mu_k\mu_m}$ for $m \neq k$ and $Cov(A_{jkt}, \bar{A}_{jk}^{-t}) = Cov(\bar{A}_{jk}^{-t}, \bar{A}_{jk}^{-t}) = \sigma_{\mu_k}^2 + \left(\sigma_{\theta_k}^2 + \frac{\sigma_\epsilon^2}{n_k}\right)/t{-}1$. The vector of reliability weights $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_K)$ can be re-expressed as:

$$\boldsymbol{\psi} = (\boldsymbol{T} + \boldsymbol{V})^{-1}\boldsymbol{T} \tag{1.19}$$

where

$$\boldsymbol{T} = \begin{pmatrix} \sigma_{\mu_1}^2 & \cdots & \sigma_{\mu_1\mu_K} \\ \vdots & \ddots & \vdots \\ \sigma_{\mu_K\mu_1} & \cdots & \sigma_{\mu_K}^2 \end{pmatrix}$$

and $\boldsymbol{V}$ is a diagonal matrix in which the k-th element is equal to $\left(\sigma_{\theta_k}^2 + \frac{\sigma_\epsilon^2}{n_k}\right)/t{-}1$. Equation 1.19 coincides with equation (3.57) of Raudenbush and Bryk (2002)'s hierarchical model, under the assumption of constant class sizes and expected value of zero for their level-2 fixed effects.

Scenario (iii): Scenarios (i) and (ii) together

The third scenario, which combines scenarios (i) and (ii), yields reliability weight for teacher VA on type-k students equal to

$$\boldsymbol{\psi}_k = \frac{\sigma_{\mu_k}^2}{\sigma_{\mu_k}^2 + \left(\sigma_{\theta_k}^2 + \sigma_\epsilon^2/n_k\right)/t - 1}. \tag{1.20}$$

This formula is discussed in Chetty et al. (2014a) as one of their special cases, and it coincides with equation (5) in Kane and Staiger (2008).

1.3 Data

1.3.1 Description of datasets

We use de-identified Chicago Public School administrative data from 2008–09 to 2016–17 school years. These data contain test scores, student demographics, attendance, suspensions, and transcripts.

Test scores. During the period of analysis, two different exams were offered, namely, the Illinois Standards Achievement Test (ISAT) and the Northwest Evaluation Association Measures (NWEA). ISAT was administered until 2013–14 to students in grades 3–8 in public schools. This test measured achievement in science (for grades 4 and 7), math, and reading. Since 2012–13 school year, the NWEA has been administered to students in grades 2–8 to assess math and reading skills. The first NWEA exam was in the fall of 2012 to have baseline test scores, and since then, the exam has been conducted every year in the spring. We treat the baseline 2012 fall exam as 2011–2012 test scores. ISAT and NWEA overlap in three school years (2011–2012, 2012–13 and 2013–14), but we select NWEA for these years because it has a larger number of test takers.

We use math and reading test scores as our main outcomes. They are normalized to have mean zero and standard deviation one by grade-subject-school year.

Student demographics. Data include race/ethnicity, gender, age, reduced and free lunch status, and special education status. The race/ethnicity variable in our data is classified into mutually exclusive categories of Black or African American, White, Hispanic, Native American/Alaskan Native, Asian/Pacific Islander, Pacific Islander/Hawaiian, and Multiracial.[3]

Student attendance and suspensions. Attendance data show the number of days a student attended and missed school. Suspensions records show the students who were suspended, type of suspension (out of school or in school), and number of days suspended.

Teacher demographics. Personnel data include teachers' race/ethnicity, gender, age, and

3. Parents report their children's race/ethnicity every year in school applications.

teaching background. We use teacher's race in secondary analyses and age for robustness checks.

Link of students to teachers. We use the transcript data to link students to their teachers. These data provide detailed course-taking information for each student, including a list of courses in which the student enrolled, their grades, and ID of the teachers who provided each grade. We identified the main math and reading courses of each student and the respective teachers who assigned the grades. Our link procedure has two shortcomings. First, we do not have actual classroom rosters of students and their teachers. Instead, we define a classroom as the group of students in the same school who are linked to the same math or reading teacher. Teachers, as a result, have one classroom per year in a subject-school cell in our constructed dataset. The second shortcoming of our linking procedure is that the teacher who entered the grade may not be the student's regular classroom teacher. We suspect these cases are relatively few.

1.3.2 Sample selection

As in Chetty et al. (2014a), we organize the data with one row per student per subject (math or reading) per school year and perform similar sample restrictions as other teacher VA papers. First, we restrict the sample to students in grades 3–8, where prior test scores are available.[4] We are left with 2.5 million math and reading test scores. Second, when a teacher teaches in multiple schools in a year, we keep the links for one school and drop the links of the other schools, as our quasi-experiment requires one teacher per school (4.4 percent of observations). Third, we remove classrooms that have more than 50 percent of students with special needs since these classes may have special teaching arrangements (5.2 percent of the remaining observations). Fourth, we drop teachers whose classrooms have less than 7 black and 7 non-black students, as we require at least 7 observations to estimate

4. During the school years 2008–09 to 2011–12, when ISAT was in place, third graders are excluded because they do not have prior test scores.

teacher VA on at least one race. We also drop teachers linked to more than 200 students because they could be mis-linked or they are entering grades for students who are not in their classrooms (1.8 percent of observations were in a small or large classroom).

The remaining data have 2.3 million student-subject-year observations. This is the core sample used in the quasi-experiment. For teacher VA estimates, we further restrict sample to students with information on prior and current test scores, demographics, and teacher assignments. This leaves us with 1.7 million student-subject-year observations.

1.3.3 Descriptive statistics of the sample

Table 1.1 reports summary statistics of the sample used to estimate teacher VA. Panel A shows student characteristics. The first row presents information on class size and number of classrooms. The average class size is 36, which is larger than typical classrooms because of our linking procedure. The second row in panel A shows the number of unique students. There are 326,750 unique students, with an average of 5.4 subject-school years. The mean test score is 0.11, higher than 0, because we normalize the test scores in the full population, which includes special education students. Half of participants are female and the average age is 10.7. CPS serves mostly minority students. The percentage of black and Hispanic students are 36 percent and 49 percent, respectively. A high percentage of students are eligible for free or reduced price lunch (86 percent) and 9 percent have special education needs. Only 1 percent of students are repeaters.

Panel B of Table 1.1 shows the characteristics of teachers. The average number of years we observe a teacher-subject (e.g., math teacher) is 3. The majority of teachers are female (84 percent). Twenty-nine percent of teacher are African-American and 19 percent are Hispanic. We calculate how many teachers never had classrooms with more than 7 black or non-black students. As we will explain below, we picked 7 to have reliable estimates of teacher effects by student race. Half of the teachers never had any classrooms with 7 or more black students and 34 percent never had any classrooms with 7 or more non-black students. These numbers

Table 1.1: Summary Statistics

	Mean (1)	S.D. (2)	N (3)
Panel A: Student characteristics			
Class size	36.04	[25.22]	48,696
Number of subject-school years per student	5.37	[2.92]	326,750
Test scores (SD)	0.11	[0.93]	1,754,812
Female	0.5	[0.50]	1,754,812
Age	10.65	[1.67]	1,754,812
Black	0.36	[0.48]	1,754,812
Hispanic	0.49	[0.50]	1,754,812
Free or reduced lunch eligible	0.86	[0.35]	1,754,812
Special education	0.09	[0.29]	1,754,812
Repeating grade	0.01	[0.11]	1,754,812
Panel B: Teacher characteristics			
Number of subject-school years per teacher	3.04	[2.07]	16,041
Female	0.84	[0.37]	15,946
Age	35.39	[10.24]	15,935
Black	0.29	[0.46]	15,946
Hispanic	0.19	[0.39]	15,946
Never had a classroom with >= 7 black students	0.51	[0.50]	16,041
Never had a classroom with >= 7 non-black students	0.34	[0.47]	16,041
Never had a classroom with >= 7 female students	0.03	[0.17]	16,041
Never had a classroom with >= 7 male students	0.03	[0.17]	16,041

Notes: Data come from de-identified administrative data of Chicago Public Schools. Sample is restricted to students with current and prior test scores, demographics, and teacher assignment. This sample is used to estimate teacher VA. First column shows the mean, second column standard deviation, and third column number of observations. For panel A, number of observations (column 3) is the number of classrooms in the first row, number of students in the second row, and number of student-subject-year observation in all other rows. For panel B, the number of observations is equal to the number of unique teachers in the data.

indicate that, for a large proportion of teachers, traditional VA models would estimate their VA on specific student subgroups and would not capture their VA on a different student population. The last two rows of Table 1.1 show the percentage of teachers who never taught female or male students. Contrary to race, virtually all teachers have a classroom with 7 or more female and male students. We will use student's sex as a validation of our multivariate model.

1.4 Characterization of Teacher VA

1.4.1 Estimation of teacher VA on black and non-black students

We describe the steps we took to estimate VA model with multivariate effects. Student type is based on race, so the two types are black and non-black. Because the majority of students in CPS are either black or Hispanic, non-black students are mostly Hispanic or brown students. We estimate teacher VA for blacks and non-blacks, separately by grade level (elementary and middle grades) and subject (math and reading).

As described in Section 1.2.2, we estimate teacher VA in student subgroups first by residualizing test scores with respect to covariates. Our set of covariates X_{ikt} is similar to that Chetty et al. (2014a) and includes: (i) cubic polynomials in prior test scores in math and reading, interacted with grade; (ii) student's sex, race/ethnicity, age, reduced and free lunch status, and special education status; (ii) logarithm of the baseline number of absences, in-school suspension and out-of-school suspension (log of the variable plus one), and grade repetition indicator; (iii) grade and year fixed effects; (iv) class- and school-year means of student characteristics; (v) cubics in class- and school-grade mean prior test scores; and (v) class size and proportion of students who are female and black.

Second, each β_k is estimated under separate subsamples using within-teacher variation. We include teacher fixed effects in the regressions. Including or excluding teacher fixed effects does not have major impact on residuals because the correlation of residuals across these two

specifications is 0.99. We compute the residuals only for classrooms with 7 or more students of type k. Classrooms with less than this number of students of this type are dropped from the VA calculations. We do this to avoid the influence of potential outliers. Fourth, we calculate class-averages of residuals by student type and form vectors of prior mean residual scores.

Fifth, we compute the variance-covariance matrix of class residuals for blacks and non-blacks across years, which is the matrix $\mathbf{\Gamma}$ in our model. Table 1.2 shows this matrix by subject-grade. Columns 1 through 4 are for math test scores, and columns 5 through 8 for reading scores. Panel A shows test scores in elementary schools and panel B for middle schools. Focusing on math, column 1 reports the auto-covariance and auto-correlation (brackets) vectors of VA on black students. This is $\sigma_{\mu_{black}s}$. Column 2 shows the cross-covariance with VA black as the leading variable and VA as lagged variable. This is $\sigma_{\mu_{black}s\mu_{nonblack}}$. Column 3 reports cross-covariance of VA on non-black as leading variable and VA black as lagged variable, $\sigma_{\mu_{black}\mu_{nonblack}s}$. Column 4 shows its auto-covariance vector, $\sigma_{\mu_{nonblack}s}$.

We can observe in panel A of Table 1.2 that the correlations decrease but not always monotonically. Another observation is that the auto-covariance vector has larger values than the cross-covariance—that is, teacher VA is more correlated with itself across years than with the other-type VA. However, the cross-correlations are relatively large, meaning that using information from one type of VA helps predict the other type.

Jumping to panels C and D of Table 1.2, they present estimates of the within-year variances and covariances $\sigma^2_{\mu_{black}}$, $\sigma^2_{\mu_{nonblack}}$, $\sigma^2_{\theta_{black}}$, $\sigma^2_{\theta_{nonblack}}$, $\sigma^2_{\varepsilon_{black}}$, $\sigma^2_{\varepsilon_{nonblack}}$, $\sigma_{\mu_{black}\mu_{nonblack}}$, and $\sigma_{\theta_{black}\theta_{nonblack}}$. Given that we only observe one classroom per teacher-year, we cannot separate $\sigma^2_{\mu_{black}}$ from $\sigma^2_{\theta_{black}}$, $\sigma^2_{\mu_{nonblack}}$ from $\sigma^2_{\theta_{nonblack}}$, and $\sigma_{\theta_{black}\theta_{nonblack}}$ from $\sigma_{\mu_{black}\mu_{nonblack}}$.[5] However, we can indirectly estimate these parameters using their lagged auto-covariances and

5. Studies that estimate teacher VA on separate subsamples and correlate teacher VA on one group with that of another group would confound the within-year cross-covariance of teacher effects with within-year cross-covariance of classroom shocks. Therefore, they over-estimate the relationship across VAs.

Table 1.2: Variance-Covariance Matrix of Teacher VA on Blacks and Non-Blacks

	Math				Reading			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Elementary schools								
Var 1:	VA black		VA non-black		VA black		VA non-black	
Var 2:	VA black	VA non-black	VA black	VA non-black	VA black	VA non-black	VA black	VA non-black
Lag of var 2								
1	0.038	0.039	0.040	0.044	0.020	0.010	0.013	0.012
	[0.425]	[0.416]	[0.426]	[0.588]	[0.337]	[0.217]	[0.24]	[0.36]
2	0.030	0.030	0.030	0.040	0.019	0.008	0.007	0.011
	[0.328]	[0.331]	[0.341]	[0.527]	[0.314]	[0.173]	[0.131]	[0.347]
3	0.028	0.023	0.033	0.037	0.018	0.008	0.007	0.008
	[0.299]	[0.24]	[0.344]	[0.494]	[0.29]	[0.159]	[0.114]	[0.266]
4	0.029	0.033	0.035	0.033	0.012	0.013	0.001	0.007
	[0.345]	[0.321]	[0.418]	[0.467]	[0.212]	[0.257]	[0.029]	[0.231]
5	0.023	0.031	0.030	0.032	0.012	0.013	0.006	0.006
	[0.292]	[0.352]	[0.351]	[0.477]	[0.214]	[0.266]	[0.133]	[0.185]
6	0.022	0.037	0.033	0.034	0.010	0.009	0.004	0.008
	[0.272]	[0.383]	[0.364]	[0.484]	[0.198]	[0.162]	[0.081]	[0.242]
7	0.019	0.062	0.020	0.033	0.019	0.029	0.009	0.005
	[0.26]	[0.479]	[0.352]	[0.452]	[0.288]	[0.456]	[0.206]	[0.181]
Panel B. Middle schools								
Var 1:	VA black		VA non-black		VA black		VA non-black	
Var 2:	VA black	VA non-black	VA black	VA non-black	VA black	VA non-black	VA black	VA non-black
Lag of var 2								
1	0.014	0.015	0.013	0.016	0.008	0.003	0.003	0.005
	[0.339]	[0.354]	[0.336]	[0.52]	[0.282]	[0.109]	[0.095]	[0.275]
2	0.009	0.014	0.011	0.013	0.006	0.004	0.004	0.005
	[0.223]	[0.326]	[0.287]	[0.427]	[0.216]	[0.143]	[0.159]	[0.244]
3	0.006	0.014	0.013	0.013	0.005	0.003	0.004	0.004
	[0.149]	[0.323]	[0.324]	[0.444]	[0.191]	[0.111]	[0.158]	[0.21]
4	0.009	0.013	0.012	0.013	0.005	0.005	0.002	0.004
	[0.217]	[0.323]	[0.312]	[0.418]	[0.178]	[0.18]	[0.075]	[0.232]
5	0.005	0.017	0.014	0.013	0.006	0.002	0.005	0.003
	[0.124]	[0.401]	[0.345]	[0.425]	[0.216]	[0.058]	[0.187]	[0.152]
6	0.010	0.016	0.015	0.013	0.007	0.006	0.007	0.004
	[0.226]	[0.352]	[0.346]	[0.395]	[0.251]	[0.216]	[0.263]	[0.254]
7	0.009	0.025	0.014	0.015	0.012	0.006	0.000	0.003
	[0.221]	[0.289]	[0.373]	[0.435]	[0.312]	[0.225]	[0]	[0.229]

Panel C: Within-year variance and covariance components for elementary schools

	VA black	VA non-black	VA black	VA non-black
Total SD	0.359	0.282	0.336	0.238
$\sigma_{\bar{\varepsilon}_k}$	0.281	0.216	0.287	0.214
$\sigma^2_{\mu_k} + \sigma^2_{\theta_k}$	0.077	0.066	0.049	0.025
σ_{μ_k} (estimated)	0.213	0.223	0.153	0.120
$\sigma_{\mu_k\mu_m} + \sigma_{\theta_k\theta_m}$	0.058		0.023	
$\sigma_{\mu_k\mu_m}$ (estimated)	0.035		0.009	

Panel D: Within-year variance and covariance components for middle schools

	VA black	VA non-black	VA black	VA non-black
Total SD	0.262	0.215	0.282	0.235
$\sigma_{\bar{\varepsilon}_k}$	0.225	0.186	0.257	0.221
$\sigma^2_{\mu_k} + \sigma^2_{\theta_k}$	0.037	0.028	0.024	0.014
σ_{μ_k} (estimated)	0.158	0.137	0.114	0.076
$\sigma_{\mu_k\mu_m} + \sigma_{\theta_k\theta_m}$	0.025		0.011	
$\sigma_{\mu_k\mu_m}$ (estimated)	0.013		0.003	

Notes: Panels A and B of this table present the variance-covariance matrix of teacher VA on blacks and non-blacks by subject and grade level. Covariance and correlation (in brackets) are reported. Panel C and D report estimates of the variance and covariance components of the parameters in the model.

cross-covariances.

To estimate $\sigma^2_{\mu_{black}}$ we use the sequence of auto-covariances $\sigma_{\mu_{black}s}$ and similarly for $\sigma^2_{\mu_{nonblack}}$ with $\sigma_{\mu_{nonblack}s}$. Using a quadratic regression of the auto-covariances on lag time, we find that the true variance of math teacher VA on black students is 0.21 and the variance of math teacher VA on non-black students is 0.23 for math in elementary schools (see panel C of Table 1.2).

We estimate $\sigma_{\mu_{black}\mu_{nonblack}}$ using the vector of cross-covariances $\sigma_{\mu_{black}\mu_{nonblack}s}$ and $\sigma_{\mu_{black}s\mu_{nonblack}}$. We find the that within-year cross-covariance, $\sigma_{\mu_{black}\mu_{nonblack}}$, is 0.035 for math in elementary schools. Converting this value in correlation terms yields a value of $\rho = 0.73$ $(= 0.035/(0.213 \times 0.223))$, which indicates that teacher VAs are highly correlated for math in elementary schools. For reading in elementary schools $\rho = 0.60$, also highly correlated. The estimated correlations for math and reading in middle schools are 0.51 and 0.38, respectively, slower than the correlations in elementary schools.

Given that the number of observations to estimate the auto- and cross-covariances diminishes the longer the lag, we set the covariances for longer periods to the values of lag 4. This is the draft limit that Chetty et al. (2014a) set to 7 in their 10-year long dataset, but we set a lower value for having fewer years of data points. Appendix Table A.3.1 shows the number of observations used to construct the variance-covariance matrix of Table 1.2.

With estimates of the variance-covariance matrix, $\boldsymbol{\Gamma}$, and covariance matrix, $\boldsymbol{\gamma}$, we construct the reliability weights $\boldsymbol{\psi}$. These weights take into account correlations within and between teacher VAs. Because of this, we are able to project estimates of teacher VA on blacks to the distribution of VA on non-blacks and vice versa. We are also able to estimate the profile of VAs for teachers with some missing information and estimate both type VAs for teachers who only taught to one student subgroup. We adjust the estimates for classroom size, n_k and number of years a teacher appear in the data.

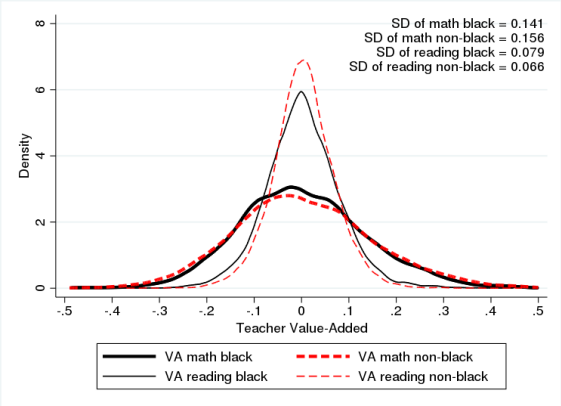
We make leave-one-year-out forecasts of teacher VA for each student type, $\hat{\boldsymbol{\mu}}_{jt}$, with equation 1.14. The next section characterizes the distributions of teacher VA.

1.4.2 Marginal distributions of teacher VA

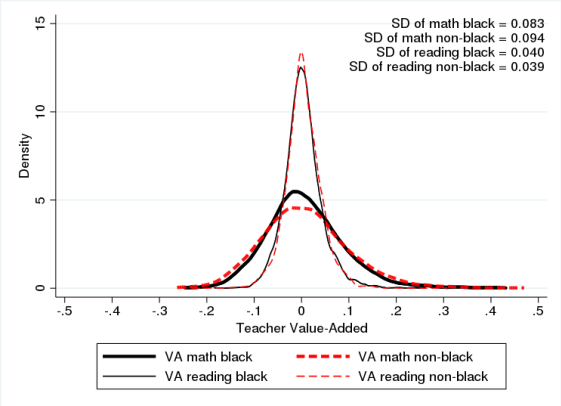
The marginal distributions of teacher VA on blacks and non-blacks by subject and grade level are in Figure 1.2. By construction, all marginal distributions have mean zero conditional on time and student type. Panel 1.2a shows the distributions for elementary schools and panel 1.2b for middle schools. In both grade levels, the math scores VA has higher variance than reading score VA. This is consistent with previous studies that also find more variability in math VA than English VA. For elementary grades specifically, there are some differences in the variance across race. The variance for math VA on black students is smaller than for non-black students, but the variance of reading VA on blacks is larger than non-blacks (however, we do not formally test this). The are no racial differences in VA for middle schools.

Figure 1.2: Empirical Distributions of VA on Black and Non-Black Students

(a) Elementary Schools



(b) Middle Schools



Notes: These figures show the empirical distribution of estimated math VA on blacks (thick dark solid line), reading VA on blacks (thin dark solid line), math VA on non-blacks (thick red dotted line), and reading VA on non-blacks (thin red dotted line). Panel 1.2a shows distributions for elementary schools and panel 1.2b for middle schools.

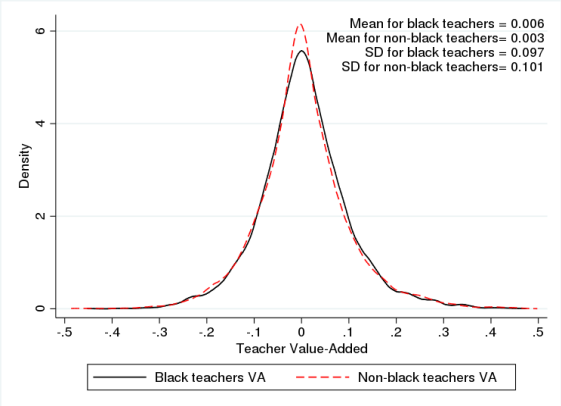
Our VA model allows us to examine differences in teacher VA on black and non-black students by teacher's race. This is related to the racial-match literature, which finds positive match effects on various outcomes when students are assigned to teachers with same demographics. Many theories exist why racial-match effects happen, including shared cultural

understanding, role model, and even teachers being more effective with same-race students (for a review, see Redding, 2019). Our approach would directly speak to racial math due to teacher effectiveness.

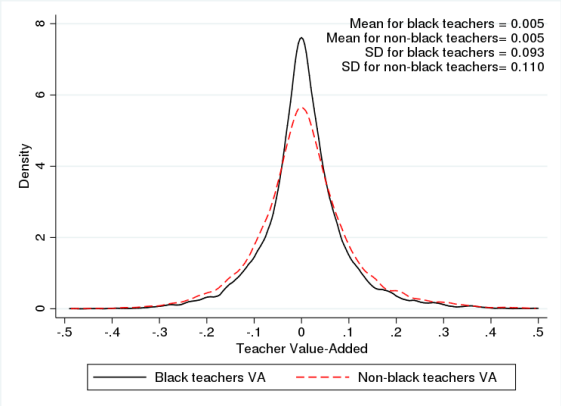
Figure 1.3 shows the empirical distribution of teacher VA on black students (panel 1.3a) and non-black students (panel 1.3b) by teacher's race. Note that the average effect by teacher's race must not necessarily be 0 because we did not use this variable in VA estimations. We do not find significant differences in the *means* of the distributions, which indicates that black and non-black teachers are on average equally effective with students, whether they share or not same race. We find that the variance of non-black teachers' VA for non-black students is slightly larger than that of black teachers.

Figure 1.3: Empirical Distributions of VA on Black and Non-Black Students by Teacher's Race

(a) VA on black students



(b) VA on non-black students



Notes: These figures show the empirical distribution of estimated VA on black students (panel 1.3a) and non-black students (panel 1.3b) by teacher's race. Black teachers' VA is the black, solid line and non-black teachers' VA is the red, dotted line.

To reconcile our findings with those studies on racial-match effects, we first reproduce their methodology and present results in Appendix Section A.2. Following Egalite et al. (2015), we employ a student fixed-effect model to estimate the relationship between student-teacher race-matching and student achievement. Our results mirror other studies in that students score higher when they are matched to teachers with similar racial characteristics.

Why do we observe positive racial match effects but no differences in VA by teacher's race? A possible explanation is differences in the source of variation used to estimate these effects. In one case we employ student fixed effects, while in the other we use teacher fixed effects and out-of-sample predictions. More investigation of these differences is required.

1.4.3 Joint distribution of teacher VA

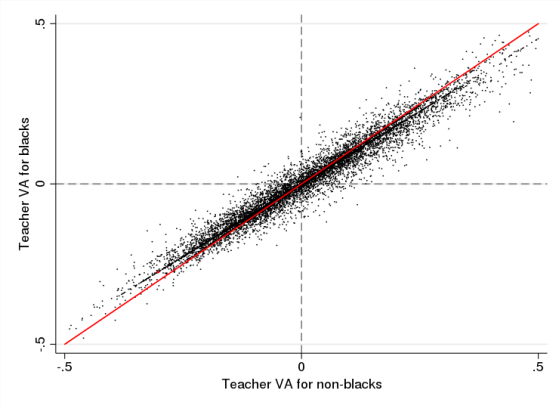
This section provides our first (observational) evidence of the existence of multivariate teacher effects. Figure 1.10 presents scatter plots of teacher VAs for math elementary schools (panel 1.4a), reading elementary schools (panel 1.4b), math middle schools (panel 1.4c), and reading middle schools (panel 1.4d). In every figure we plot a 45-degree line in red and a vertical and horizontal line that divides the graph in four quadrants at the coordinate $(0, 0)$.

Five patterns emerge from Figure 1.10. First, VAs are highly correlated, with the largest correlations for math scores in elementary and middle grades, and the lowest one for reading middle schools. The mass concentration is on the 45-degree line and close to the coordinate (0,0). Second, despite of the high correlations, we observe large heterogeneities since many points fall outside the 45-degree line. Third, the reading scores have an "X," which could indicate that the VA model is imputing many values. We will investigate more why we observe this behavior. Fourth, a large number of dots fall above the 45-degree line that refer to teachers with a comparative advantage at teaching black students. Teachers below the 45-degree line have a comparative advantage with non-black students.

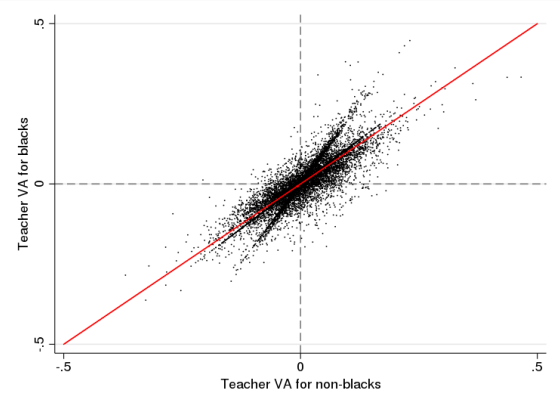
Fifth, we find observations in all four quadrants of each graph. In the first quadrant we find teachers who are above average across both dimensions; they are good teachers for both student groups. Observations in the bottom-right quadrant are teachers effective with non-black students but ineffective with black students (relative to the average teacher). The bottom-left quadrant indicates teacher below average on both dimensions, and the top-left quadrant indicates the teachers who are effective with black students but are ineffective with non-black students. Appendix Figure A.3.1 provides additional evidence of the het-

Figure 1.4: Joint Distribution of Teacher VA on Black and Non-Black Students by Subject and Grade

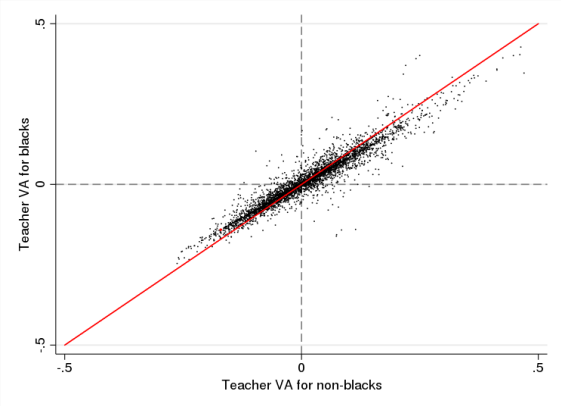
(a) Math elementary schools



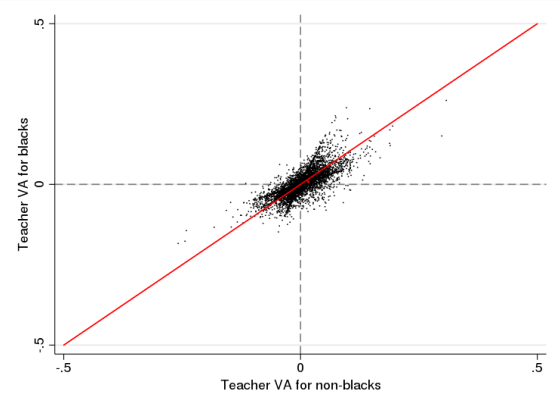
(b) Reading elementary schools



(c) Math middle schools



(d) Reading middle schools



Notes: Scatter plot of Teacher VA on blacks and non-blacks. Red line is the 45 degree line. The correlations of the unshrunken teacher effects, based on the estimated covariances and variances presented in Table 1.2 panels C and D, are 0.73, 0.60, 0.51 and 0.38 for panels 1.4a, 1.4b, 1.4c, and 1.4d, respectively.

erogeneities in teacher VA by plotting the distribution of VA on blacks by deciles of VA on non-blacks.

1.4.4 Out-of-sample forecasts

Our second (observational) evidence of multivariate teacher VA consists of predicting student test scores with own-type and other-type VA. We estimate the following regression:

$$A_{ikt} = \beta_0 + \lambda \hat{\mu}_{jkt} + \phi \hat{\mu}_{jk't} + \varepsilon_{ikt} \tag{1.21}$$

where A_{ikt} is residual student test scores, $\hat{\mu}_{jkt}$ is own-type teacher VA, and $\hat{\mu}_{jk't}$ is other-type teacher VA. Black student's own-type VA is teacher VA on black students, and similarly for non-black students. In this regression we control for math and middle school indicators as well as their interaction because VA is separately estimated by subject and grade level. Standard errors are clustered by school-cohort.

Table 1.3 presents estimates of equation 1.21. Column 1 reproduces column 1 of Table 3 in Chetty et al. (2014a) and uses own-type VA as the only explanatory variable. Column 2 uses other-type VA as explanatory variable, and column 3 includes both VA types. The number of observations is 1.1 million, lower than 1.7 million in Table 1.1 because the leave-one-out estimates of teacher VA are obtained for teachers with at least two years of data. In column 1 of Table 1.3, the coefficient 1 is statistically equal to 1, as expected, because $\hat{\mu}_{jkt}$ is the best linear predictor of A_{ikt}. Also as expected, the coefficient of other-type VA is close to 1 because of the high correlation between black and non-black VAs. However, we can reject that it is equal to 1 with high precision ($p < 0.01$). When we include both VA types, the coefficient of own-type VA close to one while that of other-type VA is close to 0.

We reproduce Figure 2 of Chetty et al. (2014a) in our Figure 1.5. Panel 1.5a refers to column 1 of Table 1.3 and panel 1.5b to column 3. Most points fall in the regression line in specification with own-type VA. The regression line is flat in the specification that

also includes other-type VA, indicating that this variable has no explanatory power once own-type VA is controlled for. These correlations are not necessarily causal because other determinants may be driving this relationship. For example, high-achieving students are sorting to schools with high-VA teachers. We will provide quasi-experimental evidence that student test scores are affected only by their specific-type VA and not by the other-type VA. In other words, there exist multivariate teacher effects that are type-specifics.

Table 1.3: Estimates of Forecast Bias

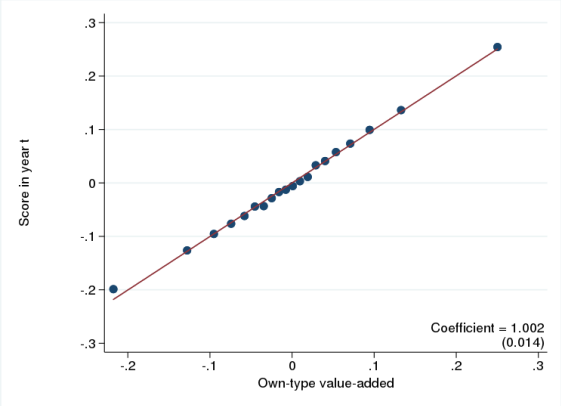
	Scores		
	(1)	(2)	(3)
Own-type VA	1.002		0.98
	(0.014)		(0.032)
Other-type VA		1.038	0.026
		(0.015)	(0.034)
p-value own-type = 1	0.89		0.52
p-value other-type = 1		0.01	0
N	1,125,545	1,125,545	1,125,545

Notes: This table reports OLS estimates of residual student test scores on estimated teacher VA. Standard errors are clustered by school-cohort and reported in parentheses. We test whether the coefficients are equal to 1 and report the resulting p-value. Own-type VA is teacher VA on blacks for blacks students and similarly for non-blacks students. Other-type VA is teacher VA on non-blacks for black students and vice versa for non-black students.

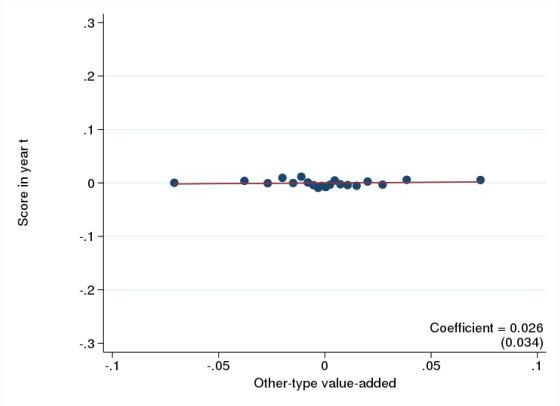
In sum, this section provides observational evidence that teacher effects are type specific. We find large heterogeneities across VA types. We also find that own-type VA explains students' test scores than other-type VA. In fact, the coefficient of own-type VA is close to 1, while the coefficient for other-type VA goes to zero once we control for own-type Va. We turn to the experimental evidence in the next section.

Figure 1.5: Effects of Teacher VA on Actual Test Scores

(a) Own-type VA



(b) Other-type VA



Notes: These figures are binned scatter plots of test scores in year t dividing sample into 20 equal-sized bins based on VA estimates. The sample consists of the students included in VA estimations. Panel 1.5a refers to column 1 of Table 1.3 and panel 1.5b to column 3. Own-type VA is teacher VA on blacks for blacks students and similarly for non-blacks students. Other-type VA is teacher VA on non-blacks for black students and vice versa for non-black students. The red lines show the best linear fit.

1.5 Quasi-Experimental Evidence

1.5.1 Empirical strategy

Following Chetty et al. (2014a), Rothstein (2017), and Bacher-Hicks et al. (2014), we exploit teacher turnover as quasi-experiment to identify the causal effects of teacher quality on student's achievement. For complete explanation see Chetty et al. (2014a). In short, when a low-VA teacher is replaced by a high-VA teacher, the change in student test scores should be equal to the change in VA. For instance, if the new teacher is 0.3σ more effective than the other teacher, the new cohort of students, who are taught by the new teacher, should score 0.3σ higher than the old cohort of students, who were taught by a lower effective teacher. Because we do not have information on classroom assignment and our definition of classroom is tied to the teacher, we cannot identify the previous cohort of students when the new teacher entered.[6] As a result, we aggregate both test scores and teacher quality

6. There are several scenarios in which the previous year's cohort of students cannot be identified—for example, when teachers in a school are re-assigned to classrooms after a new teacher enters, or when two new teachers replace two other teachers.

measures at the school-grade level. Continuing with the above example, if the school has three classrooms in a grade, the increase in grade-level average of test scores after the teacher switches should be equal to 0.1σ $(= 0.3\sigma/3)$.

We extend teacher switching quasi-experiment to allow for multivariate teacher effects and estimate the following model:

$$\Delta A_{ksgt} = \alpha + \lambda \Delta Q_{ksgt} + \phi \Delta Q_{k'sgt} + \Delta \varepsilon_{ksgt} \tag{1.22}$$

where A_{ksgt} is the average score in grade g in school s for students of type k in year t. $\Delta A_{ksgt} = A_{ksgt} - A_{ksg,t-1}$ is the across-cohort (or across-year) change in student test scores. For this quasi-experimental exercise, the estimated teacher VA is $\hat{\mu}_{jkt}^{-\{t,t-1\}}$, which is the prediction leaving out years t and $t-1$. Because we only use prior information to predict the future, these estimates are only formed for teachers with at least three years in the data. Q_{ksgt} is the type-k-student-weighted mean of $\hat{\mu}_{jkt}^{-\{t,t-1\}}$ across teachers in school s in grade g, and ΔQ_{ksgt} is the across cohort change. By leaving two years out, changes in predictions are due to changes in teacher quality rather than changes in actual student quality. Our model additionally includes $\Delta Q_{k'sgt}$, the change in other-type VA.

Assumption 3 (Conditional independence in teacher turnover quasi-experiment) *Changes in own-type teacher VA across cohort within a school grade are orthogonal to changes in other determinants of student achievement, conditional on changes in other-type teacher VA.*

$$\Delta Q_{ksgt} \perp \Delta \varepsilon_{ksgt} | \Delta Q_{k'sgt}. \tag{1.23}$$

The identifying Assumption 3 tells us that changes in teacher quality directly related to student type are uncorrelated with other determinants of student achievement once we control for changes in the other measure (other-type) of teacher quality. Under this assumption, one would expect $\lambda = 1$ and $\phi \neq 1$ if only own-type VA matters. This assumption would be violated by race-specific non-random sorting of students and teachers over time—

for instance, if students of a specific race follow teacher switchers who are effective with that race, or if student quality is differentially changing by race and teachers sort based on their comparative advantage. These types of race-specific sorting for both students and teachers are unlikely given the cost for parents and students to switch schools, and possibly neighborhood, were they to follow teachers and because school demographics often remain steady from one year to another. Furthermore, our quasi-experiment exploits yearly variation in teacher quality across grades and schools, thus high-frequency race-specific sorting of students seems unlikely.

1.5.2 Results of quasi-experiment

Table 1.4 shows our quasi-experimental estimates. Column 1 reproduces column 1 of Chetty et al. (2014a). As other authors have pointed out, it is remarkable that across settings and datasets the coefficient of the change in teacher VA is statistically equal to 1 ($p = 0.16$). The implied forecast bias is 8.5 percent, larger than previous studies, maybe due to smaller sample size, but we cannot reject that it is equal to 1. Panel 1.6a of Figure 1.6 shows the graphical representation. Column 2 of Table 1.4 uses changes in the average other-type as explanatory variable. We observe that the coefficient is much smaller than own-type and statistically different from 1 ($p < 0.01$). When we include both VA types, the coefficient of changes in own-type VA is equal to 1 ($p = 0.33$) and the coefficient for other-type is indistinguishable from 0. Panel 1.6b of Figure 1.6 graphically shows this results.

Next, we use as instrument the changes in teacher quality due to departures (i.e., switch across schools or exit from the school district), which are likely to be uncorrelated with high-frequency changes in student quality across years (Chetty et al., 2014a). Columns 4 through 6 in Table 1.4 show the IV estimates. The patterns are similar to the preceding estimates. We cannot reject that own-type VA has a 1-to-1 relationship with changes in test scores and other-type VA, while being correlated with student test scores, its effects become null once own-type VA is included.

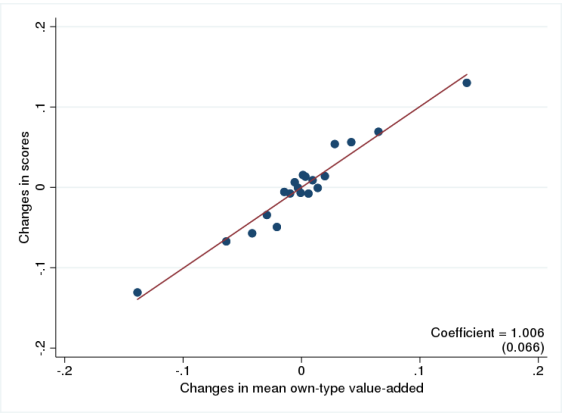
Table 1.4: Teacher-Switching Quasi-Experimental Estimates of Forecast Bias

	Changes in scores					
	(1)	(2)	(3)	(4)	(5)	(6)
Changes in mean own-type VA	1.006		1.021	0.974		0.984
	(0.066)		(0.071)	(0.123)		(0.153)
Changes in mean other-type VA		0.379	-0.029		0.698	-0.016
		(0.050)	(0.049)		(0.124)	(0.147)
p-value own-type = 1	0.92		0.76	0.84		0.92
p-value other-type = 1		0	0		0.01	0
IV teacher exit only				X	X	X
N	19,627	19,627	19,627	19,627	19,627	19,627

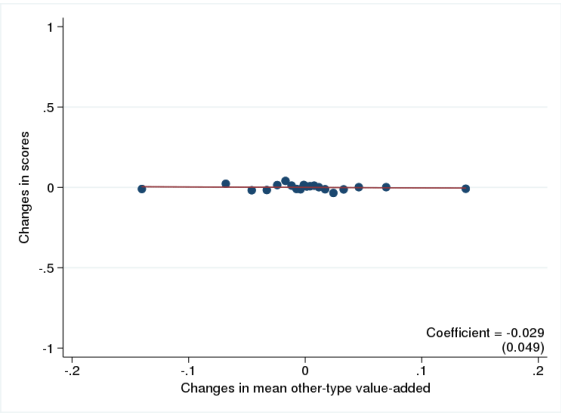
Notes: Table shows regression of changes in average test scores at the school-grade-subject level on changes in estimated VA at the school-grade-subject level. Standard errors are clustered by school-cohort and reported in parentheses. VA is estimated using two-year-out predictions that excludes times t and $t-1$. Own-type VA is teacher VA on blacks for black students and teacher VA on non-blacks for non-black students, while other-type VA is teacher VA on non-blacks for black students and teacher VA on black for non-black students We test whether the coefficients are equal to 1 and report the resulting p-value. Columns 1 through 3 show OLS regression. Columns 4 through column 6 instrument changes in mean teacher VA with changes in teacher quality when teachers exit or switch to a different school.

Figure 1.6: Teacher Switching Quasi-Experimental Evidence of Teacher Effects

(a) Own-type VA



(b) Other-type VA



Notes: These figures are binned scatter plots of changes in test scores in year t dividing the sample into 20 equal-sized bins based on changes in teacher VA estimates. The sample includes all students with information on teacher assignment. This sample is larger than sample of students used in VA estimations. Panel 1.6a refers to column 1 of Table 1.4 and panel 1.6b to column 3. Own-type VA is teacher VA on blacks for blacks students and similarly for non-blacks students. Other-type VA is teacher VA on non-blacks for black students and vice versa for non-black students. The red lines show the best linear fit.

In sum, our teacher-switching estimates support previous studies that validate VA measures. We further use this strategy to test for multivariate teacher effects and find evidence that only own-type VA affects student scores. The following section provides additional evidence of multivariate teacher VA.

1.6 Evidence Using Holdout Strategy

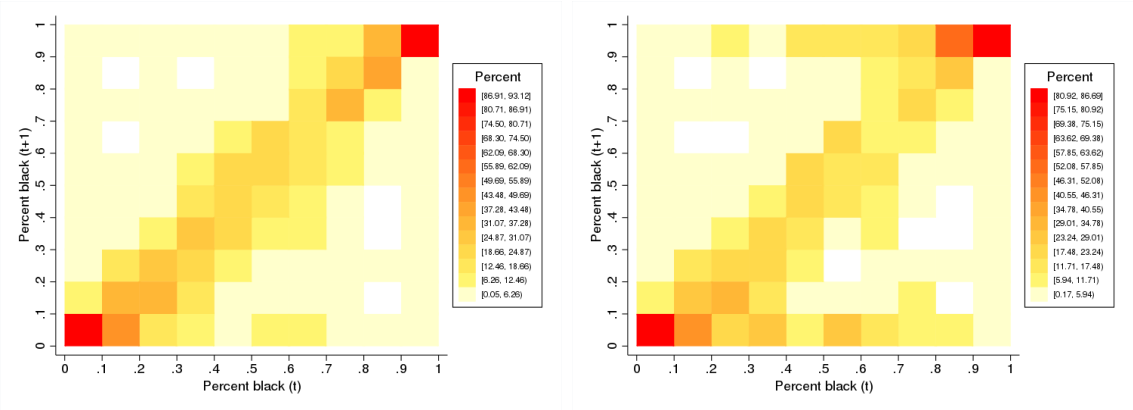
1.6.1 Holdout strategy

We continue exploring teacher turnover by examining the types of classrooms they move to after switching. Figure 1.7 shows heat maps of the conditional transition matrix (or conditional probability) of moving to a type-k classroom in year $t+1$ given that he or she currently teaches a type k' classroom. For this exercise we define classroom as the percentage of black students in the classroom. Panel 1.7a uses the sample of all switchers across grades within school or across schools while panel 1.7b restricts to switchers across schools only. Most of the moves are between classrooms of similar type, as can be shown for the dark area around the 45-degree line. When we look at switchers that move to another school, a "Z" shape appears, meaning that some teachers move to classrooms with either very few or a large proportion of black students.

Our additional test of multivariate effects works as follows. If teacher effects were homogeneous, a VA model that assumes homogeneity would accurately predict test scores when a teacher moves. If teacher effects were heterogeneous, the forecast bias of this model would be larger since the predictions would be off when teacher makes large changes in the type of classrooms. For example, if a teacher has only taught to non-black students, her VA would be on this population. If she moves to a classroom with a large proportion of black students, VA models that assume homogeneity would project her VA on non-black students, which could not be the same as her VA for black students. In other words, the prediction of homogeneous VA model would be off for teachers who make drastic moves.

Figure 1.7: Conditional Transition Matrix of Teacher Switchers Given Classroom Composition in Time t

(a) All swtichers (b) Switchers across schools



Notes: Heat map of teacher mobility. Each figure shows the proportion of teachers who move to a classroom with various percentages in $t+1$ conditional on the percentage of black students at time t. Panel 1.7a restricts sample to teachers switching from one grade to another within a school or moving between schools. Panel 1.7b restricts sample to teachers switching between schools. The x-axis is the percentage of black students in the classroom at time t, with each bin having a width of 10 percent. The y-axis is the percentage of black students at time $t+1$, after the move. Given a value of X, the sum of proportions across Y is equal to 1.

In order to isolate any correlations between test scores and teacher quality when making predictions for switchers, we split the data into two: the training and test data. Drastic switchers will be part of the *test* data, which is the holdout sample. With the *training* data, which only consists of non-movers and movers with no drastic changes in their type of classrooms, we estimate the reliability weights of the multivariate VA model. Then, we use these weights to make predictions in the *test* data. We do this exercise using both our multivariate VA model as well as the homogeneous VA model of Chetty et al. (2014a). If teacher effects were heterogeneous, a model that allows for multivariate effects would do better than a model that assumes homogeneity.

We conduct holdout sample predictions for three different subsamples of movers: (i) all switchers who move across school, (ii) switchers who move across schools *and* the percentage of black students in their classrooms *increase* by more than 20 percentage points, and, similarly, (iii) across-school switchers whose classrooms saw a *decrease* in the percentage of

black students by 20 percentage points or larger. We reproduce Table 1.3 and show estimates for the test samples.

1.6.2 Results of holdout strategy

Columns 1 through 3 of Table 1.5 report estimates under homogeneous teacher VA and columns 4 through 9 under multivariate effects. Starting with columns 1 and 4, we observe that we cannot reject the the coefficients are equal to 1 ($p = 0.09$ and 0.06 for homogeneity and multivariate VA models, respectively). Once we further restrict the holdout sample to switchers whose classrooms had an increase of black students, the coefficient under homogeneity increases to 1.30 and is statistically different from 1 at $p = 0.02$ (see column 2). On the other hand, the coefficient for multivariate teacher VA is indistinguishable from 1 $p = 0.13$. Restricting the holdout sample to switchers with a decrease in black students (i.e., increase in non-black students) increases estimates, making them statistically different from zero ($p =< 0.01$) for both cases (see columns 3 and 6). The higher estimates of homogeneous VA model's estimates may indicate that VA is being underestimated for teacher switchers.

Table 1.5: Estimates of Forecast Bias Using Holdout Samples

	Homogeneity			Heterogeneity					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Own-type VA	1.034	1.300	1.371	0.835	1.163	1.303	0.618	0.872	1.346
	(0.020)	(0.127)	(0.071)	(0.087)	(0.109)	(0.060)	(0.065)	(0.116)	(0.087)
Other-type VA							0.396	0.365	-0.055
							(0.056)	(0.158)	(0.100)
p-value own-type = 1	0.086	0.019	0	0.058	0.133	0	0	0.270	0.000
p-value other-type = 1							0	0	0
Sample	Switch school	Drastic switchers to blacks	Drastic switchers to non-blacks	Switch school	Drastic switchers to blacks	Drastic switchers to non-blacks	Switch school	Drastic switchers to blacks	Drastic switchers to non-blacks
N	379231	23291	37712	379231	23291	37712	379231	23291	37712

Notes: Table presents OLS estimate of residual student test scores on estimates of teacher VA, using holdout sample. Standard errors are clustered by school-cohort and reported in parentheses. Holdout sample is excluded from estimates of reliability weights. With these estimate we then predict teacher VA with the holdout sample only. Columns 1, 4, and 7 use the sample of switchers across schools as the holdout sample. Columns 2, 5, and 8 further restrict to movers who had an increase of at least 20 p.p. in the percentage of black students. Columns 3, 6, and 9 restrict holdout sample to switchers with a decrease in the percentage of black students.

Columns 7 to 9 repeat estimates but includes other-type VA as additional explanatory

variable. Across specifications the coefficient for own-type VA is closer to one than the coefficient of other-type VA, which in turn is closer to 0. Again, this is some suggestive evidence that own-type VA matters for student's learning more so than other-type VA.

1.7 Validation of Methodology and Robustness Check

In this section, we run an additional validation and robustness checks. A concern of estimating teacher VA by student's race is that a non-trivial amount of teachers never taught to black (or non-black) students, so their VAs on non-black (black) students are parametrically estimated. We reproduce our quasi-experimental findings using student's sex instead of race, and present results in Table 1.6. Because every teacher taught to both male and female students, the number of imputed VAs is low.

Columns 1 through 6 of Table 1.6 support the findings that teacher effects are student specifics. We cannot reject that the coefficient of changes in own-type VA is equal to 1 across specifications. The coefficient of changes in other-type VA, on the contrary, becomes different from 1 and statistically indistinguishable from 0 (see columns 3 and 6).

Another concern about our multivariate VA is the assumption that novice and experienced teachers have the same variance-covariance effects. As a robustness check, we reproduce quasi-experimental estimates, restricting sample to experienced teachers. Given that we only observe years of experience in CPS and not total years of teaching experience, we use age as a better proxy of total experience. We restrict the sample to teachers who are 30 years old or older, who are likely to have more than 3 years of teaching experience.[7]. Table 1.7 shows robustness check with this sample. Results are qualitatively similar in that own-type VA explains changes in test scores.

7. Tenure is given to teachers with more than 3 years of experience.

Table 1.6: Teacher-Switching Quasi-Experimental Estimates Using Students' Sex

	Changes in scores					
	(1)	(2)	(3)	(4)	(5)	(6)
Changes in mean own-type VA	1.014		0.996	0.966		0.777
	(0.068)		(0.118)	(0.122)		(0.285)
Changes in mean other-type VA		0.964	0.018		0.953	0.192
		(0.068)	(0.119)		(0.121)	(0.273)
p-value own-type = 1	0.84		0.97	0.78		0.43
p-value other-type = 1		0.59	0		0.7	0
IV teacher exit only				X	X	X
N	25,176	25,176	25,176	25,176	25,176	25,176

Notes: Table shows regression of changes in average test scores at the school-grade-subject level on changes in estimated VA at the school-grade-subject level. Standard errors are clustered by school-cohort and reported in parentheses. VA is estimated using two-year-out predictions that excludes times t and $t-1$. Own-type VA is teacher VA on females for female students and teacher VA on males for male students, while other-type VA is teacher VA on males for female students and teacher VA on females for male students. We test whether the coefficients are equal to 1 and report the resulting p-value. Columns 1 through 3 show OLS regression. Columns 4 through instruments with changes in teacher quality when teachers exit or switch to a different school.

Table 1.7: Teacher-Switching Quasi-Experimental Estimates Restricting Sample to Experienced Teachers

	Changes in scores					
	(1)	(2)	(3)	(4)	(5)	(6)
Changes in mean own-type VA	1.08		1.119	0.957		0.918
	(0.075)		(0.087)	(0.133)		(0.173)
Changes in mean other-type VA		0.34	-0.074		0.706	0.065
		(0.053)	(0.059)		(0.128)	(0.162)
p-value own-type = 1	0.29		0.17	0.75		0.63
p-value other-type = 1		0	0		0.02	0
IV teacher exit only				X	X	X
N	16,846	16,846	16,846	16,846	16,846	16,846

Notes: Table shows regression of changes in average test scores at the school-grade-subject level on changes in estimated VA at the school-grade-subject level. Standard errors are clustered by school-cohort and reported in parentheses. Sample is restricted to teachers who are 30 years old or older. VA is estimated using two-year-out predictions that excludes times t and $t-1$. Own-type VA is teacher VA on blacks for black students and teacher VA on non-blacks for non-black students, while other-type VA is teacher VA on non-blacks for black students and teacher VA on black for non-black students We test whether the coefficients are equal to 1 and report the resulting p-value. Columns 1 through 3 show OLS regression. Columns 4 through column 6 instrument changes in mean teacher VA with changes in teacher quality when teachers exit or switch to a different school.

1.8 Teacher comparative advantage

1.8.1 Characterization of teacher comparative advantage

Results in previous sections provide evidence that teacher effects are multivariate. In this section, we characterize teachers' comparative advantage and relates it with educational mobility. We define teacher's comparative advantage on black students as VA on blacks minus VA on non-blacks. Positive values of this variable means that teachers are more effective at teaching black students than non-black students, and negative values mean the comparative advantage is for non-black students. We will use this measure in a policy simulation that will match teachers to students based on teacher comparative advantage. The larger the comparative advantage (in absolute terms) the farther is the teacher from the 45-degree line in Figure 1.10.

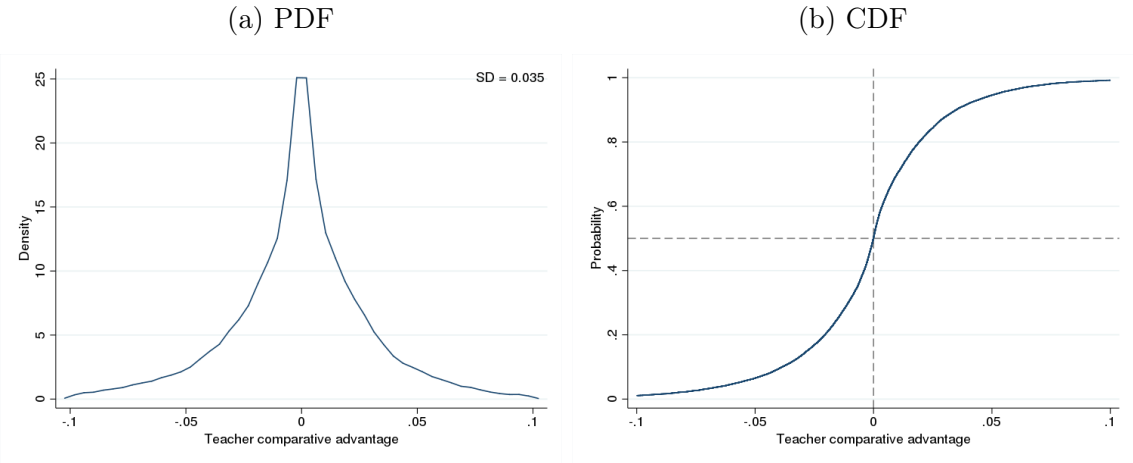
Figure 1.8 plots the probability density (panel 1.8a) and cumulative distribution (panel 1.8b) functions. The mean, standard deviation, and skewness are 0 (by construction), 0.036 and 0.395, respectively.[8] The distribution is approximately symmetric and highly concentrated at 0. We reject the Shapiro-Francia test for normality. The 25th percentile is -0.013σ and the 75th percentile is 0.013σ. About 90 percent of observations are between -0.05 and 0.05 as can be seen in the CDF. To put these numbers into perspective, the racial-match effects in our data is 0.009σ (see Appendix Table A.3.2, column 7). Matching teachers based on comparative advantage may have larger effects than matching based on race.

1.8.2 Relationship of comparative advantage and educational mobility

Are students being assigned to teachers with comparative advantage for their own type? We document the extent of student-teacher match based on teacher's comparative advantage

8. The mean and standard deviation of teacher's comparative advantage on black students for math in elementary schools are -0.001 and 0.032, respectively; for reading in elementary schools, these values are -0.002 and 0.040; for math in middle schools they are -0.001 and 0.032; and for reading and middle schools, they are -0.001 and 0.034.

Figure 1.8: Probability Density and Cumulative Distribution Functions of Teacher Comparative Advantage on Black Students

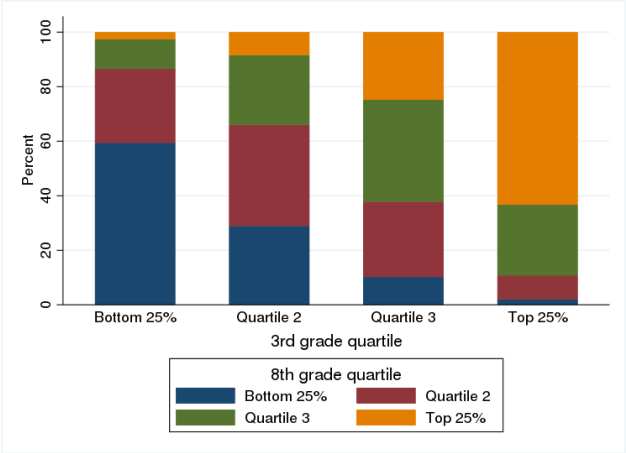


Notes: A teacher's comparative advantage is defined as the difference between her VA on black students minus her VA on non-black students. A positive value means that she is more effective on black students than non-black students and viceversa.

using the concept of educational mobility. In the same spirit of income mobility, we define educational mobility as the change in the test score distribution between the end and start of schooling. Because we have data from 4th to 8th grade, we use lagged baseline test scores in 4th grade—that is, 3rd grade test scores—as the starting point and end-of-year test scores in 8th grade as ending point. We divide the sample into quartiles based on 3rd- and 8th-grade test score distributions and present their joint distribution in Figure 1.9. This graph is also a transition matrix. Each x-value indicates the quartile in 3rd grade and each color refers to a different quartile in 8th grade. As shown in other studies, we observe low mobility (or SES disparities). 60 percent of students who started at the bottom quartile stayed at that quartile in 8th grade while 3 percent of them ended up in the top quartile. Conversely, 63 percent of students at the top quartile ended up in the same quartile in 8th grade while only 2 percent transitioned to the bottom quartile. For the next exercise we focus on students in the bottom quartile in 3rd grade.

We classify a student as high achiever if he or she moved up from the bottom to the top quartile ("Bot-Top") and as low achiever if stayed in the bottom quartile ("Bot-Bot").

Figure 1.9: Proportion of Students in Test Score Quantiles in 3rd and 8th Grades



Notes: This figure depicts the joint distribution of test score quartiles in 8th grade by quartiles of 3rd grade test scores.

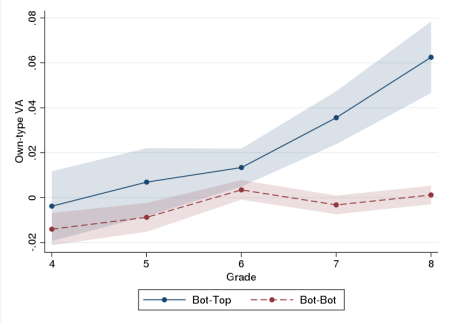
Panel 1.10a of Figure 1.10 plots the profile of own-type VA by mobility status.[9] There is a mechanical correlation between *levels* in teacher VA and mobility status, which is shown by the positive trend for high achievers and a flatter one for low achievers.

Educational mobility may not necessarily have a mechanical correlation with *differences* in teacher VA. High-achieving students may be assigned to high-VA teachers who may not have comparative advantage for their types. Panel 1.10b plots the profiles of teacher comparative advantage by mobility status. We find that high-achieving students are being assigned to more effective teachers based on their type, and the matches improve the more they progress through school. On the other hand, lower-achieving students are not being matched with teachers with comparative advantage. In 4th grade they are assigned to less effective teachers (comparative advantage is negative), but this negative matching disappears in later grades. We point out that the correlation between comparative advantage and educational mobility may not causal for several reasons. For instance, high-SES parents are better at identifying good teachers for their children than lower-SES parents, or high-VA teachers may sort to more advantaged schools because of personal preferences.

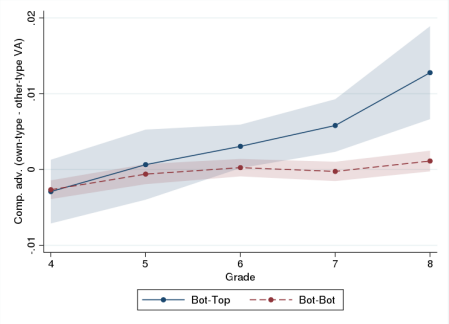
9. We keep the first test score if a student repeats a grade.

Figure 1.10: VA Profiles for High-Achieving and Low-Achieving Students

(a) Own-Type VA



(b) Teacher's Comparative Advantage



Notes: High-achieving students ("Bot-Top") are those who started in the bottom quartile of the test score distribution in 3rd grade (first grade of data) and moved up to the top quartile in 8th grade (last grade of data). Low-achieving students ("Bot-Bot") are those in started in the bottom and stayed at that quartile in the last grade. Panel 1.10a plots the own-type VA profile from 4th to 8th grade. Because teacher VA and our definition of educational mobility are mechanically correlated, there is a positive trend in the VA for high-achieving students and a flatter trend for low-achieving students. Panel 1.10b plots teacher's comparative advantage profile, which is not mechanically correlated with educational mobility. Many factors can explain why high-achieving students are being matched with teachers with comparative advantage—for instance, highly-effective teachers sort to schools with high-achieving students. This finding is, therefore, descriptive and not causal.

1.9 Policy Implications

Ignoring multivariate effects when promoting or de-selecting teachers may have negative implications. Prior studies suggest laying off teachers in the bottom of the VA distribution. However, if we assume homogeneous teacher effects and de-select teachers in the bottom 5th (10th) percentile, 33 percent (31 percent) of them would otherwise be above this percentile under multivariate effects (i.e., either black or non-black estimated VA would fall above this percentile in their own type-specific distribution).

For future work, we plan to conduct a policy simulation that matches teachers to classrooms based on teachers' comparative advantage. If teachers were paid bonuses equal to their marginal productivity and teacher labor markets functioned as free markets, the market equilibrium would be efficient. For example, suppose that teachers are paid a bonus equal to \$100 per 1σ increase in classroom average test scores. Assume also that there are two teachers, A and B. Teacher A increases test scores of black students by 0.3σ and scores

of non-black students by 0.2σ. Teacher B has an effect of 0.4σ on black students and 0.1σ on non-black students. Assume there are two classrooms, X and Y, with different classroom compositions. Classroom X has 60 percent black and 40 percent non-black students and classroom Y has 40 percent black and 60 percent non-black students. Teacher A's bonus payment if she teaches in classroom X would be \$26 $(= 100 * (0.3 * 0.6 + 0.2 * 0.4))$ and her bonus for classroom Y would be \$24 $(= 100 * (0.3 * 0.4 + 0.2 * 0.6))$. Teacher B's bonus payments would equal to \$28 and \$22 for classrooms X and Y, respectively. In this hypothetical scenario, teacher A would be willing to pay up to \$2 more to teach in classroom X rather than in Y, while teacher B would bid up to \$6 more to be in classroom X. The equilibrium assignment is thus teacher A with classroom Y and teacher B with classroom X.

One achieves the same matching equilibrium if a central planner uses teachers' comparative advantage to assign teachers to classrooms. Teacher A's bid to be in classroom X is equal to her relative contribution on that classroom—that is, it is equal to her comparative advantage on black students times the difference between classrooms in the proportion of black students ($\$2 = 100 * (0.3 - 0.2) * (0.6 - 0.4)$)—and similarly for teacher B's bid ($\$6 = 100 * (0.4 - 0.1) * (0.6 - 0.4)$). The resulting match is thus teacher B with classroom X and teacher A with classroom Y. The central planner's maximization problem would to maximize student learning measured by test scores. The constraint faced by the central planner will be how far a teacher can move to a different school. So, for instance, a teacher can be moved only to schools within the same neighborhood she actually teaches or within a 1-mile radius of her current school.

Student-teacher matching based on *average* VA could yield different results. For instance, teacher A could be assigned to classroom X if teacher A were of race black and the central planner would match her with the classroom with the largest proportion of black students. With our estimated multivariate VA and comparative advantage, we can simulate various matching mechanisms and compare their efficiency and equity impacts.

1.10 Conclusions

Are teacher effects student specific? We addressed this question by first proposing and developing a flexible multivariate VA model that relaxes homogeneity assumption and allows teacher effects to vary by student's race and drift over time. Using CPS data, we estimate teacher VA on black and non-black students and characterize their distributions. Using observational, quasi-experimental and holdout methods, we find evidence that teacher effects vary by student's race and, thus, some teachers have comparative advantage on student black students and others on non-black students. Our results also indicate that students are affected by their own-type teacher VA. That is, black students are affected by their teachers' estimated VA on black students rather than the estimated VA on non-black students, and vice versa.

Our work speaks to the fairness of teacher quality measures. Assuming homogeneous teacher effects when in fact they are multivariate can have important implications, especially when schools are segregated and teachers serve different student populations. If one student subgroup is more disadvantaged than other groups, teachers who serve that group would be penalized by traditional VA models, even though these models control for several student characteristics. A potential solution is to employ a flexible VA model with teacher effects that vary by a policy-relevant student characteristic, such as race. This way, teachers' VAs can be projected onto the same VA distribution to make more comparable assessments.

Our work also speaks to the way teachers are matched to schools and classrooms. With multivariate teacher effects, one could employ the concept of comparative advantage to improve these matches. Can we achieve a more efficient, and possibly more racially equitable, outcome by reallocating teachers to classrooms based on their comparative advantage? We plan to address this question in future work.

CHAPTER 2

EVALUATOR BIAS: UNPACKING THE RELATIONSHIP BETWEEN CLASSROOM CHARACTERISTICS AND CLASSROOM OBSERVATION RATINGS

Co-authors: Lauren Sartain and Andrew Zou

2.1 Introduction

Amid concerns that low-quality teachers were not being identified for remediation or removal by many districts across the country, the federal Race to the Top (RTTT) grant competition incentivized many states to reform their teacher evaluation policies. Most of the nation's largest districts have implemented new evaluation systems that evaluate teachers based on a well-defined classroom observation rubric and student growth metrics (Steinberg and Donaldson, 2016). One goal of this major overhaul was that teachers would be held accountable for their classroom performance, enabling districts to identify high-performing teachers for rewards or promotions; it would also create a clear path to exit persistently low-performing teachers. In order to use evaluations for high-stakes career decisions in particular, the teachers who are being evaluated—as well as the general public—need assurance that the evaluation systems are generally prioritizing the things that teachers do that most directly relate to improved student outcomes and learning. Further, there should be certainty that teacher performance is measured accurately.

Regarding measurement, a lot of public attention has been paid to value-added measures (VAMs), which are often included in evaluation systems as a way to quantify a teacher's unique contribution to student learning during the school year. While measurement issues with VAMs have been well documented (Rothstein, 2009, 2017), most teachers do not teach in tested subjects and grade levels, so VAMs cannot be estimated for most teachers. In addition, classroom observation ratings almost always receive the heaviest weight in the calculation of

evaluation ratings for all teachers, even those who have VAMs. For these reasons, this chapter focuses on the measurement of teacher practice via classroom observations. The classroom observations that comprise evaluations are most often conducted by school administrators who know teachers using a rubric that delineates levels of performance in multiple areas of classroom instruction and management.

There is certainly room for error in the classroom observation ratings. Teachers raise concerns about the extent to which evaluations do capture their actual performance. For example, in Chicago, the setting of this chapter, teachers have reported the evaluations can be influenced by the students in the classroom. Another concern is that teachers are being evaluated by school leaders who do not have a background in their content area or grade level. Or that it is more difficult to get "distinguished" ratings when you serve more disadvantaged populations (e.g., classrooms with higher proportions of special education students or high rates of students with behavioral challenges). Finally, the observation rubric itself may not be culturally competent and therefore it would be less appropriate to evaluate some student populations. At the root of concerns like these is that the application of the observation rubric used in the evaluation system may not result in the accurate assessment of teacher performance, or at least an assessment of teacher performance that is independent of classroom context. These kinds of arguments, if true, suggest that teachers working with less advantaged student populations could systematically receive lower ratings than their practice warrants.

In this chapter, we empirically test the extent to which the students in teachers' classrooms influence their evaluator-assigned classroom observation ratings. We answer the following research questions: (i) To what extent are various measures of teacher quality (i.e., administrator-assigned classroom observation ratings, student survey responses of classroom experiences) influenced by the characteristics of students in the classroom? (ii) Would adjusting classroom observation ratings for student characteristics result in a different ranking of teachers than unadjusted classroom observation ratings? We address these questions using

administrative data of Chicago Public Schools (CPS).

Our contribution to the literature is fourfold. First, the ratings in our setting are authentic in that school administrators assign them following an in-person classroom observation, while other studies have used data collected by video and rated by trained observers. Therefore, the findings in this paper are more generalizable for policy and practice. Second, our data consists of the universe of CPS teachers, the third largest school district in the U.S.; therefore, our findings have more external validity than previous experimental studies. Third, we are able to use five years of observation data, tracking teachers over time as they experience different classroom contexts. Specifically, with the longitudinal nature of the data, in this chapter we employ fixed-effect methods to estimate how year-to-year changes in the characteristics of teachers' classrooms correlates with ratings. Finally, we are able to compare evaluator-assigned classroom observation ratings—which are part of the high-stakes evaluation system—to student reports of experiences with teachers, which are not part of the evaluation. This comparison allows us to test if the evaluator and potentially the rubric itself are more sensitive to changes in classroom composition than are students, which has implications for teacher evaluation policies.

Our results indicate that having higher-achieving students, less disruptive students, and fewer disadvantaged students, even when the teacher remains in the same school, is associated with better observations. On the other hand, we find no statistically significant associations of classroom characteristics with student ratings. If year-to-year variation in classroom composition is idiosyncratic, these findings could be interpreted as causal. We then simulate the distribution of teacher performance under a hypothetical policy that adjusts observation ratings for classroom characteristics, more akin to value-added measures. 30 percent (8.6 percent) of teachers who would otherwise be in the bottom 5th (10th) percentile of the unadjusted rating distribution are above that percentile on the adjusted distribution.

The remainder of the chapter is organized as follows. We provide a literature review on teacher quality. Then, we describe the teacher evaluation system in CPS. Next, we describe

the data, sample, and methodology for estimating the effects of classroom characteristics on measures of teacher quality. Results are presented in the following section. Next, we discuss the implications of a policy that adjusts observation ratings. We conclude in the final section.

2.2 Prior Literature

Past research has found that high-quality teachers can have a significant effect on student achievement and later life outcomes (Sanders and Horn, 1998; Rockoff, 2004; Rivkin et al., 2005; Aaronson et al., 2007; Chetty et al., 2014b). Studies have also shown that both subjective and objective evaluations of teacher effectiveness can predict future student achievement (Jacob and Lefgren, 2008; Rockoff and Speroni, 2011). Strong teaching can be even more important for lower-achieving students (Aaronson et al., 2007), suggesting that personnel decisions ranging from hiring to classroom staffing to training are critical for a school district's ability to reduce student achievement gaps. Given the extensive literature supporting the importance of teachers, policymakers have placed an increased focus on developing effective teacher evaluation methods. In particular, the Obama administration's Race to the Top (RTTT) grant competition encouraged states to commit to making teacher evaluations more rigorous and more objective. These reformed evaluation systems aimed to achieve two goals: 1) identify low-performing teachers for remediation or removal, and 2) provide all teachers with concrete feedback about how to improve teaching practice.

In recent years, education initiatives like RTTT have also shifted teacher evaluation systems to include multiple measures of teacher practice. For example, 27 states required teacher ratings to be based on multiple measures of student growth and achievement, and 44 states required classroom observations to be incorporated into teacher evaluations (Doherty and Jacobs, 2013). However, there remains differences in how each state and school district chooses to use evaluation policies, especially in nontested grades and subjects (Steinberg and Donaldson, 2016). Many school districts have used measures of student growth on test scores

(e.g., value-added measures) as a component of their teacher evaluation systems. While some argue that value-added measures provide unbiased estimates of teachers' impact on student achievement (Kane and Staiger, 2008; Chetty et al., 2014a), others express concern over their validity and reliability (Rothstein, 2009, 2017; Papay, 2011). But in reality, most teachers do not receive value-added scores because they do not teach subjects or grade levels that standardized tests measure. In CPS during 2016–17, only 27 percent of classroom teachers taught in tested subjects and grade levels. Recent teacher evaluation reforms have encouraged the use of evidence-based rubrics to guide classroom observations, aiming to make principal evaluations more objective and to provide a clear picture of what distinguished instruction looks like. The classroom observation component often receives the highest weight of all measures of teaching practice when constructing a final evaluation rating. Ultimately, these classroom observations and the evaluator-teacher conferences that accompany them are at the core of teacher evaluation reform, despite the focus on value-added measures.

With the redesign of teacher evaluation systems, some research has focused on "validating" classroom observation ratings by comparing them to value-added measures. A study using findings from Chicago's Excellence in Teaching Pilot found that, on average, teachers with higher classroom observation ratings had significantly higher value-added measures (Sartain et al., 2011). Students also showed the most growth in classrooms of highly rated teachers, and the least growth in classrooms of poorly rated teachers. These findings are supported by the five observation instruments used in the Measures of Effective Teaching (MET) project, which were positively associated with student achievement gains (Kane and Staiger, 2012). However, there are also concerns that teachers' observation ratings may be influenced by things beyond their control, some of which are explicitly controlled for in value-added model specifications. For example, it may be more difficult to demonstrate mastery of certain classroom practices if a teacher's students are below grade level, highly disruptive, or frequently absent from school. Certainly, an unintended and undesirable consequence of evaluation reforms would be that teachers receive low observation ratings because of their

classroom or school context rather than because of the quality of their classroom practice. For these reasons, it becomes increasingly important to understand how potential biases can affect the reliability and validity of observation scores.

A prior study on CPS' teacher evaluation system found that teachers serving more disadvantaged students tended to have lower observation and value-added scores (Sporte and Jiang, 2016), though this relationship was stronger for observation scores than for value-added scores. This study did not conclude whether these discrepancies were due to differences in instructional quality or a biased evaluation system. Other studies also conclude that classroom characteristics strongly correlate with teacher evaluation scores. The research finds that factors such as low income, limited English proficiency, and low prior achievement are associated with lower observation scores in other studies, consistent with the findings from CPS (Chaplin et al., 2014; Whitehurst et al., 2014).

While these findings consistently show the same relationships, they can be interpreted in different ways. One explanation may be the nonrandom sorting of teachers into classrooms and schools. More experienced and more highly qualified teachers tend to teach higher achieving students, while newer teachers are assigned a higher proportion of minority, low-income, and poor students (Lankford et al., 2002; Borman and Kimball, 2005; Clotfelter et al., 2006a; Kalogrides and Loeb, 2013). Researchers found that administrators, teachers, and parents all play a role in the nonrandom sorting of teachers (Paufler and Amrein-Beardsley, 2014). Teacher mobility decisions can also contribute to the sorting, as much of it occurs across districts and schools rather than within schools (Boyd et al., 2011; Goldhaber et al., 2015). Nonrandom sorting, while problematic for other reasons, can help explain the correlation between classroom characteristics and teacher evaluation scores.

Another explanation is that teaching in more challenging settings may make it more difficult to receive high ratings on the classroom observation rubrics. Teachers working in high-poverty schools or with high concentrations of underserved students in their classrooms may find it more difficult to provide quality instruction to students due to the differences

in organizational support that they receive. New teachers are often working in disorganized schools and work in a less supportive environment than other teachers (Kraft and Papay, 2014). In addition, more experienced teachers improve more quickly from additional experience than newer teachers, further exacerbating the difficult teaching conditions that new teachers face (Sass et al., 2012). Thus, it could be possible that the same teacher may be more effective teaching in a classroom or school with advantaged students than one with disadvantaged students. This can further contribute to the strong relationship between classroom characteristics and teacher evaluation scores.

But we are concerned that there still remains bias in teacher evaluation systems, especially with classroom observations. Observation ratings may not only reflect information about what teachers are doing in the classroom but also about the preparation and characteristics of students in the classroom. Whitehurst et al. (2014) try to address this problem by examining the distribution of teacher observation scores by the average incoming achievement level of their students. They found that only 9 percent of teachers assigned low-achieving students scored in the top quintile of observation scores, while 37 percent of teachers assigned high-achieving students scored in the top quintile. After adjusting for classroom characteristics, the discrepancy between scores largely disappeared. This led them to conclude that this was a sign of rater bias. However, this form of adjustment does not account for the nonrandom sorting of teachers to classrooms or other forms of bias.

A pair of past studies take advantage of the MET project to explore the relationship between teacher observation scores and classroom composition (Steinberg and Garrett, 2016; Campbell and Ronfeldt, 2018). The project sought to identify teacher evaluation methods that could both provide teachers with valuable feedback and recognize effective teachers. Because the project randomly assigned teachers to students, researchers did not have to consider nonrandom sorting in their results, although nonrandom sorting of teachers into schools was still possible. Both studies hoped to exploit this randomization to gain a better understanding of the effect of classroom composition on observation scores. Steinberg and Garrett

(2016) conclude that teachers with higher achieving students performed better on measures that evaluate their ability to create a learning environment. Campbell and Ronfeldt (2018) also find a relationship between incoming classroom achievement and teacher observation ratings. The authors believe both rater bias and actual differences in instructional quality may explain their estimates, but cannot disentangle the effects of the two.

In this chapter, we seek to understand further the role of teacher's students in shaping their classroom observation ratings. While the previous literature relies on the MET data, which has many advantages, our setting is authentic in that the ratings are done by evaluators (either principals or assistant principals) who know the teachers and students. The observations are conducted in real time and in person, as opposed to the MET classrooms where the instruction was videoed and the ratings were conducted by trained external observers. We also contrast two ratings of teacher quality: 1) the administrator-assigned classroom observation ratings, which are high stakes, and 2) student reports of instructional quality on an annual survey, which is not used in the teacher evaluation system. Comparing these two reports will provide some insights into the different ways that administrators and students observe the classroom experience. Finally, we extend the previous literature by simulating the distribution of teacher evaluation ratings if classroom observations were adjusted for the students in the classroom. We also discuss the implications of using this type of adjustment.

2.3 Teacher Evaluation in Chicago Public Schools

In response to state legislation, CPS reformed its teacher evaluation system to include multiple measures of teacher practice and rubric-based classroom observations of teacher practice. The new system called Recognizing Educators Advancing Chicago (REACH) was launched district-wide in 2012–13 and was still in place as of 2018–19. CPS' goals for REACH are focused on improving student learning experiences via improved instruction, which includes providing a common definition of high-quality teaching and supporting "ongoing conversa-

tions between school leaders and teachers to encourage growth and improvement" (School, 2019). Teachers receive REACH scores on a 100–400-point scale, which is then transformed into four rating categories: Unsatisfactory, Basic, Proficient, and Distinguished. While the district's goals focus on instructional improvement, there are stakes associated with receiving Unsatisfactory or Basic ratings. Tenured teachers with these lower ratings are placed on annual evaluation schedules (rather than every other year). For tenured teachers rated Unsatisfactory, they undergo a remediation plan and are reevaluated; if progress is not made, the district can move to dismiss those teachers. Nontenured teachers with Unsatisfactory or Basic ratings cannot make progress toward tenure in those years and are subject to dismissal.

The REACH scores and ratings are based on multiple components: classroom observation ratings (70 percent of the final rating) and student growth metrics (30 percent of the final rating). The classroom observations occur four times during the evaluation period, and a certified principal or assistant principal conducts the observations and rates the quality of teacher practice. The rubric for the observations is based on the Charlotte Danielson Framework for Teaching. The Danielson Framework was initially developed as a formative tool to help teachers, particularly novice teachers, build competency in various aspects of classroom instruction. Now, many districts across the country, including CPS, use it as an evaluation tool[1]. CPS' version of the framework includes four domains of teaching practice: Planning and Preparation, Classroom Environment, Instruction, and Professional Responsibility. Across these four domains, there are 19 more specific components of practice (e.g., Managing Student Behavior, Using Assessment in Instruction). Figure 2.1 provides an example of the rubric associated with one of the Framework components.[2] Teachers receive ratings on a four-point scale in each of these 19 components, and all of these ratings across each observation are summarized in a single teacher practice score.

Student growth is calculated based on two potential measures. The first is value-added

1. There are some concerns that the Danielson Framework is being misused as an evaluation tool since its initial purpose was more formative. This chapter does not address those concerns directly.

2. See https://cps.edu/ReachStudents/Documents/FrameworkForTeaching.pdf for the complete rubric.

Figure 2.1: Example of a Rubric Associated with the Instruction Domain of REACH

Domain 3: Instruction

Component	*Unsatisfactory*	*Basic*	*Proficient*	*Distinguished*
3a: Communicating with Students • *Standards-Based Learning Objectives* • *Directions for Activities* • *Content Delivery and Clarity* • *Use of Oral and Written Language*	Teacher neither clearly communicates standards-based learning objective(s) to students nor addresses their relevance to learning. Teacher's directions and procedures are confusing to students. Teacher's explanation of content is unclear or inaccurate; explanations do not connect with students' knowledge and experience. Teacher's spoken and written language is unclear and incorrect. Vocabulary is vague, incorrect, or inappropriate for the students' ages and levels of development, leaving students confused.	Teacher does not communicate the standards-based learning objective(s) to students or does not address their relevance to learning. Teacher clarifies directions and procedures after initial student confusion. Teacher's explanation of content contains minor errors, and/or some portions are clear while other portions are difficult to follow; explanations occasionally connect with students' knowledge and experience. Teacher's spoken and written language is unclear or incorrect. Vocabulary is limited or inappropriate for the students' ages or levels of development.	Teacher clearly communicates standards-based learning objective(s) to students and addresses their relevance to learning. Teacher clearly communicates directions and procedures. Teacher's explanation of content is clear and accurate, and connects with students' knowledge and experience. Teacher's spoken and written language is clear and correct. Vocabulary is appropriate for the students' ages and levels of development.	Teacher clearly communicates standards-based learning objective(s). Teacher guides students to articulate the relevance of the objective(s) to learning. Teacher clearly explains directions and procedures, and anticipates possible student misunderstanding. Teacher's explanation of content is thorough, accurate, and clear, enabling students to develop a conceptual understanding of content while making connections to their interests, knowledge, and experience; students contribute to extending the content by explaining concepts to their classmates. Teacher's spoken and written language is expressive, and builds on students' language development and understanding of content. Vocabulary is appropriate for the students' ages and levels of development; students contribute to the correct use of academic vocabulary.

Source: CPS (2019b).

measures based on the Northwest Evaluation Association (NWEA) test for grades 2-8, so teachers in reading or math in grades 3-8 receive one of these value-added measures. For all teachers, with or without NWEA measures, the district has also developed subject- and grade level–specific assessments called performance tasks. Teachers administer these tasks to their students at the beginning and end of the year, and the same teachers also assess their students' level of performance. The district then calculates student growth based on the scores that students received in the performance tasks.

The observation component is the focus of this chapter, as there is already extensive literature about the appropriate use and construction of teacher value-added measures based on student test score data. Further, a better understanding of classroom observation ratings is important for two primary reasons. First, most teachers do not teach in tested grades or subject areas, meaning that they cannot receive traditional teacher value-added measures. Second, even for teachers who can receive value-added measures, classroom observations typically receive higher weights in the calculation of evaluation ratings. For example, in CPS for value-added teachers, 70 percent of their evaluation ratings are calculated based on classroom observations. Ultimately, classroom observation ratings typically determine the vast majority of a teacher's evaluation rating, and teachers tend to value the feedback they provide. As a result, understanding whether evaluations capture teacher's performance or classroom context is policy relevant.

2.4 Data and Methodology

2.4.1 Data sources

We use longitudinal data from CPS teachers and students from the 2011–12 to 2016–17 school years. At the teacher level, we have personnel files, which include the school of employment and teacher demographic characteristics, and REACH files (2012–13 to 2016–17). From the student-level administrative data, we have demographic information, school enrollment, as well as academic performance information. Our data also include student survey reports of classroom experiences, although these data are only available for students in grades 6–8.

Teacher-level data. The teacher-level personnel data include information on teachers' demographics, such as gender, race, and age, as well as their teaching background, such as years of experience in CPS, education level, and tenure status. The REACH teacher evaluation data set includes each teacher's final REACH score and categorical rating, as well as scores on the subcomponents of the system: classroom observation scores from the Danielson Framework, teacher-level value-added scores where available, and scores on the district-developed performance task.

Student-level data. The student enrollment data include gender, race, and free or reduced-price lunch (FRL) status, as well as the school enrolled and the student's grade level. We also have data on student behaviors and academic performance. Specifically, we have student attendance records, which include the number of days students missed each year and the total number of days enrolled. We also observe whether or not students were reported for disciplinary infractions and whether or not those infractions resulted in in- or out-of-school suspensions. For test score data, we use the Northwest Evaluation Association Measures of Academic Progress (NWEA) scores. The NWEA is a standardized assessment on math and reading for grades 2–8 that has been administered every spring since 2012–13. In the first year of implementation, this assessment was also conducted in the fall of 2012 to obtain baseline test scores. Therefore, we treat the baseline 2012 fall exam as 2011–2012 test scores

for the first year of lagged standardized test scores. For each student, we generate lagged attendance rates, suspension indicators, and test score performance.

Survey data on students' classroom experiences is taken from the CPS My Voice, My School survey. The survey is administered annually to all CPS students from grades 6 to 12. During the period of our analysis, the completion rates for the survey were around 80 percent. The survey asked students to report on their experiences with peers and teachers, attitudes, and curricular activities in school. We extract the items that specifically asked about the students' math, English, and science teachers and courses. We create an index of these course-specific survey items to correspond to the Danielson Framework domains of The Classroom Experience and Instruction. The two constructed indexes are *survey instruction* and *survey management*. (See Appendix Table B.0.1 for a list of the items used in both indexes.) We note that these reports of student experiences with teachers are not part of the teacher evaluation system.

Linking students and teachers. We use student transcript files in order to link students to teachers. These transcripts list all courses taken by each student and have an identifier for the teacher who assigned the final grades. Because the REACH data are at the teacher level, we construct classroom rosters as all students in a school who have the same teacher, irrespective of the courses and grades he or she teaches. For example, suppose that a teacher taught English and math in 5th grade: every student who had this teacher would belong to the same classroom roster. Similarly, if a teacher taught music in 6th and 7th grades, all her students would belong to the same classroom. There are two limitations of our linking procedure. First, the teacher who assigned the final grade may not be the student's regular teacher. We believe that this is not very common across schools. Second, we are pooling students who may have been in different classrooms. Since our REACH data do not specify the classrooms in which evaluators rated teachers, our linking procedure is our best guess to identify teachers' students. Classroom characteristics are based on the characteristics of the linked students.

To link student survey responses to teachers, we employ a similar procedure but with an exception: we focus on math, English, and science teachers because these are subjects asked in the surveys. For each student in transcript files, we identify his or her main math, English and science courses and assign the graders as her teachers. Here, a classroom roster is the set of students with the same math (English or science) teacher, irrespective of the grade. Therefore, if a math teacher taught in 6th and 7th grades, all her students would belong to the same classroom. Also, if a teacher taught writing and reading, which are English courses, her students would be in the same classroom roster. We use these subject-specific links to construct classroom averages of student responses.

2.4.2 Sample description

We perform the following sample restrictions. First, we include only teachers with students in grades 3–8 as data on lagged test scores, attendance rates, and suspension rates are available. Second, classrooms with fewer than seven students with baseline test scores are dropped to reduce imprecisions when computing class mean test scores. Third, we drop classrooms with more than 200 students because such students are likely to be mis-linked. After these sample restrictions, the analytic personnel data consists of 10,854 unique teachers who appear in the data 2.7 years on average; thus, the total number of teacher-year observations is 28,944. The survey subsample consists of 1,933 unique teachers who appear in the data at the same rate as the whole sample (2.7 years). The total number of teacher-year observations for this subsample is 5,188.

Table 2.1 provides summary statistics for teacher demographics, teacher evaluation scores, and classroom characteristics. The columns are split between the entire sample and only teachers with a survey score. Most teachers are female (76 percent). About half of the teachers in the district are non-Hispanic white (52 percent), 24 percent are non-Hispanic black, and 19 percent are Hispanic. On average, CPS teachers are 40 years old with about nine years of experience in CPS, and 73 percent of them are tenured. The average final

REACH score in the sample is 312 on a 400-point scale, the average observation score is 3.2 on a 4-point scale, mean performance task score is 47.3 on a 120-point scale, and the mean value add summative is 50.3, which is the estimated value-added rescaled on a 120-point scale. The domains that make up the Danielson Framework (classroom observation) have an average between 3.11 to 3.38 on a 4-point scale. Our constructed survey measures have mean zero but standard deviation lower than one because they were standardized before imposing sample restrictions. At the classroom level, average class size is 53, which is greater than the usual 20–25 students due to our linking procedure. Half of the students are male (0.51), the majority belong to minority groups (42 percent Hispanic, 35 percent Black), 85 percent have FRL status, and 18 percent are in a special education program. Classroom average of baseline test scores is close to zero and the within-classroom standard deviation of test scores is 0.77. Average attendance rate is 95 percent and suspension rate is 7 percent.

The teachers in the survey subsample have similar characteristics as the whole sample. 79 percent of teachers are female, 49 percent are white, 29 percent are black, and 15 percent are Hispanic. They have the same age, number of years of experience and tenure status as the whole sample. Their teacher quality scores are identical to the whole sample. The characteristics of their classrooms are very similar as well. In other words, the survey sample mirrors the rest of the teachers.

How do classroom observation scores correlate with the other quality measures, teacher and classroom characteristics? Table 2.2 presents these correlations, focusing on the performance score and its four domains. Panel A column 1 shows that observation scores and their domains are highly correlated, ranging from 0.78 (domain 4) to 0.97 (domain 3). The observation score is also highly correlated with the final performance score (0.95) due to its large weight in the computation. On the other hand, it has a low correlation with the performance task (0.10) and survey instruction score (0.10). The observation score is positively correlated with the value add summative (0.24), and it has a positive association with the survey management score (0.24).

Table 2.1: Summary Statistics of Main and Survey Samples

	Main Sample			Survey Sample		
	Obs. (1)	Mean (2)	Std. Dev. (3)	Obs. (4)	Mean (5)	Std. Dev. (6)
Panel A: Teacher Demographics						
Number of years per teacher	10854	2.67	1.28	1949	2.73	1.28
Female	28944	0.76	0.42	5206	0.79	0.41
White	28944	0.52	0.5	5206	0.49	0.5
Black	28944	0.24	0.42	5206	0.29	0.46
Hispanic	28944	0.19	0.39	5206	0.15	0.36
Age	28944	40.34	10.64	5206	40.4	10.34
Years of CPS experience	28944	8.95	8.02	5206	8.94	7.84
Tenure status	28944	0.73	0.45	5206	0.74	0.44
Panel B: Teacher Quality Measures						
Final score	28944	312.4	37.29	5206	311.9	36.81
Observation score	28944	3.2	0.45	5206	3.22	0.45
Performance task	26339	47.32	23.6	4775	40.74	18.74
Value add summative	17569	50.27	12.39	4036	50.14	12.95
Domain 1	28920	3.21	0.52	5202	3.22	0.51
Domain 2	28930	3.25	0.47	5202	3.26	0.47
Domain 3	28931	3.11	0.47	5202	3.13	0.47
Domain 4	28921	3.38	0.56	5202	3.4	0.55
Survey instruction score	5207	0.04	0.88	5206	0.04	0.88
Survey management score	5206	0.07	1.01	5206	0.07	1.01
Panel C: Classroom Characteristics						
Class size	28944	53.08	33.65	5206	53.91	32.81
% male	28944	0.51	0.08	5206	0.52	0.09
% Hispanic	28944	0.49	0.39	5206	0.46	0.39
% Black	28944	0.35	0.41	5206	0.4	0.42
% free lunch	28944	0.85	0.22	5206	0.87	0.19
% special ed	28944	0.18	0.2	5206	0.22	0.26
Test score (lagged)	28944	-0.03	0.55	5206	-0.09	0.59
Test score SD (lagged)	28944	0.77	0.15	5206	0.77	0.14
Attendance rate (lagged)	28944	0.95	0.02	5206	0.95	0.02
Suspension rate (lagged)	28944	0.07	0.1	5206	0.09	0.1

Notes: Data come from CPS administrative data for the 2011-12 to 2016-17 school years. Main sample is teachers in the REACH data who were linked to seven or more students with test scores and at most 200 linked students. Survey sample is teachers who taught math, reading or science in grades 6-8 and who were linked to the student survey data. Teacher quality measures are described in Section IV. Classroom characteristics are averages of student characteristics.

Table 2.2: Correlation of Teacher Quality Measures, Teacher Characteristics, and Classroom Characteristics

	Observation Score	Domain 1	Domain 2	Domain 3	Domain 4
	(1)	(2)	(3)	(4)	(5)
Panel A: Teacher Quality Measures					
Domain 1	0.924	1			
Domain 2	0.916	0.768	1		
Domain 3	0.966	0.847	0.857	1	
Domain 4	0.776	0.693	0.65	0.688	1
Final score	0.954	0.881	0.878	0.919	0.741
Performance task	0.096	0.097	0.083	0.084	0.091
Value add summative	0.24	0.206	0.244	0.231	0.168
Survey instruction score	0.099	0.076	0.119	0.095	0.056
Survey management score	0.244	0.212	0.253	0.226	0.189
Panel B: Teacher Characteristics					
Female	0.093	0.096	0.084	0.077	0.096
White	0.175	0.175	0.136	0.145	0.168
Black	-0.231	-0.223	-0.197	-0.212	-0.222
Hispanic	0.027	0.016	0.044	0.02	0.021
Age	0.002	0.001	0.01	0.007	-0.036
Years of CPS experience	0.159	0.138	0.164	0.155	0.094
Tenure status	0.235	0.216	0.227	0.229	0.15
Panel C: Classroom Characteristics					
Class size (log)	0.091	0.093	0.074	0.086	0.071
% male	-0.029	-0.023	-0.025	-0.029	-0.027
% Black	-0.332	-0.297	-0.326	-0.303	-0.292
% free lunch	-0.261	-0.241	-0.216	-0.262	-0.219
% special ed	-0.026	-0.027	-0.012	-0.029	-0.023
Test score (lagged)	0.274	0.257	0.234	0.269	0.228
Attendance rate (lagged)	0.171	0.156	0.165	0.156	0.153
Suspension rate (lagged)	-0.256	-0.227	-0.252	-0.239	-0.211

Notes: Correlation table of classroom observation score and its domains. Teacher quality measures are described in Section IV. Classroom characteristics are averages of student characteristics.

Panel B of Table 2.2 presents correlations with teacher demographics. We can observe in column 1 that a teacher's gender, Hispanic ethnicity, and age are not correlated with observation scores. Being white is positively associated with these scores (0.18), while being black is negatively correlated (-0.23). Having more years of experience is somewhat correlated with scores (0.16), and teachers with tenure have higher scores (0.23).

In Panel C of Table 2.2, we correlate the scores with classroom characteristics. Classroom observations have a low correlation with log of class size (0.09), the percentage of male students (-0.03), and percentage of students with special education (-0.03). The scores are negatively correlated with the proportion of black students (-0.33), free lunch status (-0.26), and average baseline suspension rate (-0.26). They are positively correlated with attendance rate (0.17) and test scores (0.27).

The four domains have a similar pattern as the observation score (columns 2–5). These correlations show that teachers with more advantaged students tend to score well in every dimension. These findings are consistent with previous studies (Steinberg and Garrett, 2016; Campbell and Ronfeldt, 2018; Sporte and Jiang, 2016; Chaplin et al., 2014; Whitehurst et al., 2014). However, to provide more causal evidence, we exploit within-teacher changes in classroom composition.

2.4.3 Empirical strategy

The main goal of this chapter is to understand the extent to which classroom composition influences teachers' classroom observation ratings. Ideally, the observation ratings should reflect things about the teachers' practices, classroom management skills, instructional skills, etc. However, there are concerns that the classroom observations may also capture the ways in which the students influence the classroom environment. To investigate the effect of classroom characteristics on classroom observation scores, we employ the following empirical model:

$$A_{it} = \alpha + C'_{k(i)t}\delta + X'_{it}\beta + \tau_t + \varepsilon_{ikt} \tag{2.1}$$

where i indexes teacher, k classroom, and t time. The outcome A_{it} is a measure of teacher i's performance in time t; $C_{(k(i)t)}$ is a vector of characteristics of classroom k taught by teacher i; X_{it} refers to teacher's characteristics; τ_t is year fixed effects; and ε_{ikt} is an error term. The coefficient δ tells us how variation in classroom characteristics correlates with measures of teacher quality. Model 2.1 is estimated with and without teacher characteristics. We further exploit the panel structure of our data to re-express Model 2.1 as:

$$A_{it} = \alpha + C'_{k(i)t}\delta + \gamma_i + \tau_t + \varepsilon_{ikt} \tag{2.2}$$

where γ_i is teacher fixed effects. Including teacher fixed effects will control for invariant observable and unobservable teacher's characteristics. Thus, the coefficient δ will be estimated with year-to-year variation of classroom characteristics within teachers. Our main identifying assumption requires variation in the composition of classrooms to be uncorrelated with teacher quality. The parameter of interest, δ, is the effect of classroom characteristics on measures of teacher quality.

Our main outcome is the classroom observation scores and the four domains: (1) planning and preparation, (2) classroom environment, (3) instruction, and (4) professional responsibilities. All outcomes are standardized by year to have mean zero and standard deviation one to ease interpretation of estimates. Our secondary outcomes are students' survey responses about their teachers. We standardize each index by year to have mean zero and standard deviation one as well.

Teacher characteristics include sex, race and ethnicity, tenure status, age, age squared, experience teaching in CPS, and experience squared. Age is calculated as of September of each academic year, and experience is computed as the number of years elapsed since the

teacher was hired.[3] Therefore, age and experience are not multicollinear. Some specifications of Model 2.1 further include school fixed effects so that comparisons are across teachers within the same school.

To obtain classroom characteristics, we aggregate student-level characteristics to the teacher level. We compute the percentage of male students, black students,[4] FRL eligible students, and students with special education. We also compute the classroom means of lagged attendance rate, lagged suspension rate, and lagged test scores. We use lagged values to avoid endogeneity since teacher quality may affect contemporary students' behaviors and test scores but not past performance. Student's test score is the average of his or her math and reading scores, each of which is previously normalized by year and grade to have mean zero and standard deviation one. In addition to classroom-level averages, we include the log of class size, which is the total number of students taught by the same teacher, and the standard deviation of lagged test scores within classrooms. The latter variable is intended to capture the degree of homogeneity in students' baseline skills.

Our definition of classroom is much broader than the actual classrooms in which teacher observations occurred because we pool all students who were taught by the same teacher. If a teacher teaches in more than one classroom, we are, in fact, including students who may have not been present during the evaluation. We are incorporating measurement error in the explanatory variable. If yearly changes in the classroom composition is larger in large groups than in small groups, our estimates would be attenuated toward zero. Additionally, because teachers may select the classrooms in which they are evaluated, our estimates would be biased downward if teachers select their easiest-to-teach classrooms. The estimates that we present are thus conservative.

3. If the hiring date in CPS is missing, experience is set to zero for the first year a teacher appears in the data and is increased by one for each year that teacher remains in the data. We do not observe years of experience outside CPS.

4. We omit other race and ethnic groups because the majority of students are either black or Hispanic in CPS

2.5 Results

Table 2.3 presents our primary results of the effects of classroom characteristics on teacher observation ratings. Column 1 shows estimates of Model 2.1 with classroom characteristics as explanatory variables, column 2 further controls for teacher demographics, and column 3 is column 2 with school fixed effects. All these specifications include year fixed effects. In column 1 we observe that classroom observation scores have statistically significant associations with every classroom characteristic. Teachers in large classrooms tend to score higher. The share of male students, black students, and FRL students is negatively associated with teacher ratings, while classrooms with a higher proportion of special education children have higher ratings. Teachers with higher achieving and more homogeneous students are given higher scores than teachers with lower achieving and more heterogeneous students. In addition, teacher ratings are positively associated with last-year attendance rate and negatively associated with last-year suspension rates. Findings are qualitatively the same after controlling for teacher demographics (see column 2).

In column 3, the coefficient of class size turns negative and that of suspension rate becomes positive. The relationship with the other classroom characteristics has the same direction as in the other specifications. Having higher-achieving students and fewer disadvantaged students, even when the teacher remains in the same school, is associated with better observations.

Column 4 presents the estimates of Model 2.2, which is Model 2.1 with teacher fixed effects. This is our preferred specification because it controls for invariant observable and unobservable teacher characteristics and classroom characteristics that are constant over time. It exploits yearly variation in classroom composition and controls for cases when teachers are systematically assigned to specific classrooms or student types—for instance, when a teacher is always assigned the higher achieving group. Moreover, if yearly variation in student cohorts is exogenous, our strategy estimates the causal effect of classroom characteristics on teacher ratings.

Table 2.3: Effects of Classroom Characteristics on Classroom Observation Scores

	Classroom observation scores			
	(1)	(2)	(3)	(4)
Class size (log)	0.0349***	0.0471***	-0.0340***	0.000325
	(0.0135)	(0.0131)	(0.0126)	(0.0132)
% male	-0.356***	-0.317***	-0.219***	-0.169***
	(0.0836)	(0.0806)	(0.0697)	(0.0606)
% black	-0.586***	-0.471***	-0.201*	-0.446***
	(0.0234)	(0.0275)	(0.1030)	(0.0656)
% free lunch	-0.286***	-0.279***	-0.142*	-0.0402
	(0.0498)	(0.0483)	(0.0862)	(0.0708)
% special ed	0.642***	0.667***	0.252***	0.198***
	(0.0519)	(0.0501)	(0.0477)	(0.0634)
Test scores (lagged)	0.385***	0.335***	0.227***	0.129***
	(0.0228)	(0.0221)	(0.0221)	(0.0223)
Test scores SD (lagged)	-0.326***	-0.261***	-0.0882**	-0.0498
	(0.0486)	(0.0464)	(0.0424)	(0.0371)
Attendance rate (lagged)	1.095**	1.381***	1.614***	0.506
	(0.5320)	(0.5110)	(0.4970)	(0.4630)
Suspension rate (lagged)	-0.355***	-0.275***	0.109*	0.002
	(0.0790)	(0.0752)	(0.0665)	(0.0588)
Specification	Class char. only	Teacher char.	School f.e.	Teacher f.e.
Observations	28,944	28,944	28,944	28,944
R-squared	0.171	0.236	0.414	0.861

Notes: Correlation table of classroom observation score and its domains. Teacher quality measures are described in Section IV. Classroom characteristics are averages of student characteristics.

Results in column 4 show that classroom characteristics have an effect on teacher ratings. Specifically, a 10 percentage point increase in the percentage of male students decreases teacher ratings by 0.017σ, and a 10 percentage point increase in the percentage of black students has a larger negative effect of 0.045σ. On the contrary, raising the percentage of special education students by 10 percentage points increases observation scores by 0.013σ, and increasing baseline test scores by 0.1σ raises scores by 0.02σ. All of these effects are statistically significant at the 1 percent level. The other classroom characteristics have effects indistinguishable from zero.

Using our preferred specification (Model 2.2), we estimate the effects of classroom composition on the four domains of classroom observations and present results in Table 2.4. Having more male students negatively affects domains 2–4, and having more black students negatively impacts all domains. The share of students with special needs increases the scores for domains 1–3. An increase in mean baseline test scores raises the scores in all four domains, with domains 1-3 having a statistically significant effect at the 1 percent level and domain 4 at the 10 percent level. More heterogeneous classrooms, measured by the standard deviation of baseline test scores, have lower scores in domains 2 and 4. The higher the percentage of students with FRL status, the lower the scores for domain 4.

Results from Tables 2.3 and 2.4 provide evidence that classroom observation scores and its domains are affected by the type of students a teacher has. Characteristics such as the percentage of male students, black students, and sometimes more heterogeneous students have a consistent negative effect on teaching scores, whereas the proportion of high-achieving students have a positive effect. If our identification assumption holds—that is, if yearly variation in classroom composition is independent of teacher quality—our results show that rater bias must be playing a role when rating teachers. Teachers with less advantaged students are being penalized. Remember that our estimates are lower bounds, or at least conservative, because we do not observe the exact classrooms in which the evaluation took place.

Table 2.4: Effects of Classroom Characteristics on the Four Domains of Classroom Observation Scores

	Domain 1 (1)	Domain 2 (2)	Domain 3 (3)	Domain 4 (4)
Class size (log)	-0.0081	-0.00447	0.0142	-0.0111
	(0.0147)	(0.0139)	(0.0141)	(0.0175)
% male	-0.078	-0.223***	-0.144**	-0.264***
	(0.0694)	(0.0652)	(0.0651)	(0.0867)
% black	-0.365***	-0.568***	-0.358***	-0.371***
	(0.0686)	(0.0693)	(0.0687)	(0.0778)
% free lunch	-0.0682	0.0965	-0.0249	-0.216**
	(0.0777)	(0.0769)	(0.0735)	(0.0930)
% special ed	0.171**	0.177***	0.241***	0.0634
	(0.0727)	(0.0684)	(0.0666)	(0.0902)
Test scores (lagged)	0.0984***	0.143***	0.147***	0.0563*
	(0.0246)	(0.0238)	(0.0232)	(0.0297)
Test scores SD (lagged)	-0.0143	-0.0755*	-0.014	-0.163***
	(0.0419)	(0.0402)	(0.0394)	(0.0517)
Attendance rate (lagged)	0.455	0.384	0.0554	0.964
	(0.5030)	(0.4900)	(0.4800)	(0.6050)
Suspension rate (lagged)	0.0431	-0.0426	0.00185	-0.0113
	(0.0653)	(0.0634)	(0.0613)	(0.0798)
Specification	Teacher f.e.	Teacher f.e.	Teacher f.e.	Teacher f.e.
Observations	28,920	28,930	28,931	28,921
R-squared	0.814	0.837	0.842	0.715

Notes: Domain 1 is Planning and Preparation, Domain 2 is The Classroom Environment, Domain 3 is Instruction, and Domain 4 is Professional Responsibilities. Each column is a separate regression where the dependent variables are domains 1 through 4 and the explanatory variables are classroom characteristics. All specifications include year fixed effects and teacher fixed effects. Robust standard errors in parentheses. Asterisks denote *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

We now turn to ratings given by students to investigate whether they are influenced by the classroom context. Because the surveys were administered to students in grades 6–8 and the questions were only about their math, reading, and science teachers, the sample size drops from 28,920 to 5,202. To make sure that differences in the sample is not driving the results, we reproduce column 4 of Table 2.3 and columns 1–4 of Table 2.4 with the survey subsample.

Table 2.5 shows estimates of the effects of classroom characteristics on classroom observation ratings and its four domains, restricting the sample to teachers with survey data. All specifications include teacher fixed effects. Compared to Tables 2.3 and 2.4, the standard errors of all coefficients in Table 2.5 are about three times larger, and this could have impacted the significance level of various coefficients. In spite of this, we still observe that the percentage of male and black students negatively affects several outcomes, while baseline test scores have a positive effect on scores. Other characteristics such as the percentage of special education children and average attendance rate are no longer statistically significant at conventional levels, but their signs remain the same.

Having shown that our findings in the overall sample still apply to the survey subsample, we analyze the extent to which student ratings of their teachers vary by classroom context. Table 2.6 reproduces columns 1 and 4 of our main table (Table 2.3) using survey instruction and survey management scores as dependent variables. Columns 1 and 2 have survey instruction scores as the dependent variable and columns 3 and 4 have survey management scores. Columns 1 and 3 presents associations, which is Model 2.1. Several characteristics have strong associations with the survey scores, such as percent black, free lunch, and special education status, baseline test scores, and attendance rate. In within-school comparisons shown in columns 3 and 4, there is no evidence of statistically significant effects of the variables.

In sum, our quasi-experimental strategy of exploiting within-teacher year-to-year variation in classroom characteristics shows that evaluators' scores are affected by the types of

Table 2.5: Effects of Classroom Characteristics on Classroom Observation Scores and Its Four Domains for Samples of Teachers with Survey Data

	Classroom observation scores	Domain 1	Domain 2	Domain 3	Domain 4
	(1)	(2)	(3)	(4)	(5)
Class size (log)	0.0707	0.0571	0.0369	0.0899	0.0294
	(0.0507)	(0.0524)	(0.0559)	(0.0567)	(0.0651)
% male	-0.414*	-0.176	-0.462*	-0.398	-0.478
	(0.2490)	(0.2660)	(0.2760)	(0.2720)	(0.3140)
% black	-0.435*	-0.392	-0.462*	-0.548**	0.0654
	(0.2300)	(0.2390)	(0.2510)	(0.2440)	(0.2870)
% free lunch	-0.0175	0.0712	-0.161	0.118	-0.276
	(0.2370)	(0.2540)	(0.2680)	(0.2510)	(0.2970)
% special ed	0.368	0.217	0.429	0.457*	0.0489
	(0.2400)	(0.2500)	(0.2790)	(0.2640)	(0.3070)
Test scores (lagged)	0.215***	0.132	0.269***	0.220**	0.105
	(0.0819)	(0.0902)	(0.0877)	(0.0875)	(0.1080)
Test scores SD (lagged)	0.00215	0.0266	-0.062	0.0248	-0.0713
	(0.1260)	(0.1410)	(0.1390)	(0.1340)	(0.1720)
Attendance rate (lagged)	2.495	2.435	1.847	2.186	2.334
	(1.5340)	(1.6540)	(1.7160)	(1.5450)	(1.9400)
Suspension rate (lagged)	-0.0445	0.0906	-0.274	0.0668	-0.0501
	(0.2470)	(0.2700)	(0.2740)	(0.2560)	(0.3080)
Specification	Teacher f.e.	Teacher f.e.	Teacher f.e.	Teacher f.e.	Teacher f.e.
Observations	5,202	5,202	5,202	5,202	5,202
R-squared	0.889	0.889	0.892	0.903	0.829

Notes: Sample is teachers with survey data. Domain 1 is Planning and Preparation, Domain 2 is The Classroom Environment, Domain 3 is Instruction, and Domain 4 is Professional Responsibilities. Each column is a separate regression where the dependent variables are classroom observations and domains 1 through 4, and the explanatory variables are classroom characteristics. All specifications include year fixed effects and teacher fixed effects. Robust standard errors in parentheses. Asterisks denote *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table 2.6: Effects of Classroom Characteristics on Survey Instruction and Management Scores

	Survey Instruction Scores		Survey Management Scores	
	(1)	(2)	(3)	(4)
Class size (log)	-0.0724**	-0.00851	-0.0406	-0.0322
	(0.0288)	(0.0310)	(0.0325)	(0.0382)
% male	0.274	0.00907	-0.421*	0.283
	(0.2170)	(0.1780)	(0.2370)	(0.2120)
% black	0.444***	0.118	-0.249***	0.106
	(0.0489)	(0.1850)	(0.0565)	(0.2400)
% free lunch	0.800***	0.0122	0.372***	0.0593
	(0.1140)	(0.1320)	(0.1190)	(0.2250)
% special ed	-0.244*	-0.146	0.449***	0.122
	(0.1380)	(0.2080)	(0.1460)	(0.1910)
Test scores (lagged)	0.165***	0.0336	0.359***	0.0204
	(0.0636)	(0.0475)	(0.0654)	(0.0649)
Test scores SD (lagged)	0.353***	0.13	0.112	-0.135
	(0.1080)	(0.0873)	(0.1230)	(0.1060)
Attendance rate (lagged)	9.582***	0.571	5.565***	1.356
	(1.0820)	(0.8940)	(1.1730)	(1.0800)
Suspension rate (lagged)	-0.112	-0.127	-0.297	-0.284
	(0.1980)	(0.1440)	(0.2180)	(0.1900)
Specification	Class char. only	Teacher f.e.	Class char. only	Teacher f.e.
Observations	5,206	5,206	5,206	5,206
R-squared	0.108	0.963	0.093	0.95

Notes: Sample is teachers with survey data. Survey instruction and management scores are indexes constructed based on student responses to CPS My Voice, My School survey. Each column is a separate regression where the dependent variables are survey instruction score and survey management scores and the explanatory variables are classroom characteristics. Columns 2 and 4 in addition include teacher fixed effects. All specifications include year fixed effects. Robust standard errors in parentheses. Asterisks denote *** p<0.01, ** p<0.05, * p<0.1.

students a teacher has. The less advantaged and more diverse classrooms, the lower the score the teacher receives. The exception is classrooms with a large percentage of special education students whose teachers are scored higher. On the contrary, we find that the ratings given by students are independent of classroom characteristics.

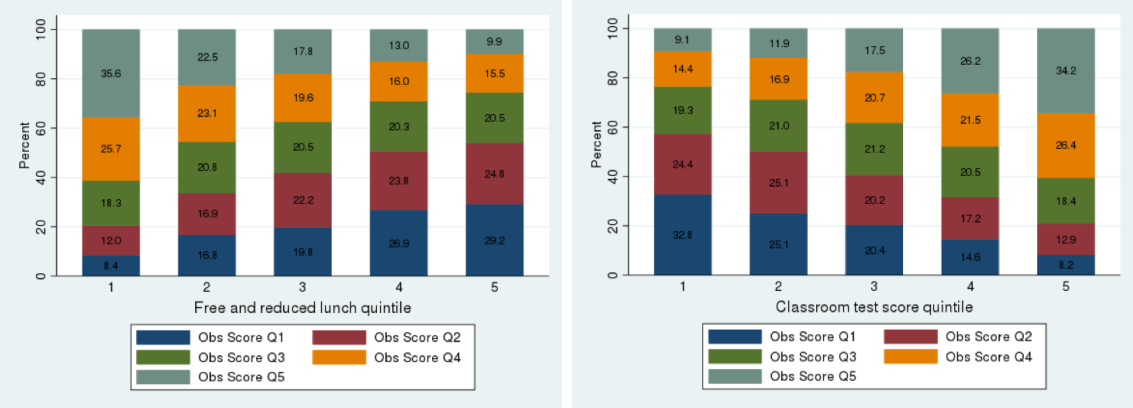
2.6 Policy Simulation: Adjusting Observation Scores for Classroom Composition

An approach to level the playing field and not penalize teachers with more disadvantaged students is to adjust observation scores by classroom characteristics. To illustrate this procedure, we plot in Figure 2.2 the joint distributions of teachers' observation scores with the percentage of FRL students (panel 2.2a) and classroom average of baseline test scores (panel 2.2b). In each figure, we divide the sample into five equal-sized groups based on the specified classroom characteristic and show the proportion of teachers who scored in the first, second, third, fourth, and fifth quintiles. Figure 2.2 shows that classrooms with lower FRL rates and higher test scores have higher proportions of teachers scoring in the top two quintiles than other classrooms. On the other hand, classrooms with higher FRL rates and lower test scores tend to have teachers whose observation scores are in the bottom two quintiles. For example, panel 2.2a of Figure 2.2 shows that 36 percent of the teachers in classrooms with the lowest rates of FRL students (FRL quintile 1) are in the top quintile of observation score distribution, while only 10 percent of teachers in the poorest classrooms (FRL quintile 5) received an observation score in the top quintile. Panel 2.2b reveals similar score differences with teachers in classrooms with varying test scores. 34 percent of teachers in the highest test performing classrooms (test score quintile 5) received a score in the top quintile, while only 9 percent in the lowest performing classrooms are in that quintile.

Researchers and school administrators have faced similar concerns when using test scores as an evaluation metric. To account for differences in classroom composition, they have

Figure 2.2: Quintiles of Unadjusted Observation Scores by Quintiles of Classroom Characteristics

(a) Free/Reduced Price Lunch Percentage (b) Baseline Test Scores



Notes: These figures show the joint distribution of classroom observation ratings with FRLP (panel 2.2a) and baseline test scores (panel 2.2b) by quaintiles.

used value-added measures that control for prior test scores, race, gender, low-income status, etc. In the case of classroom observations, Whitehurst et al. (2014) use data from four urban districts to adjust observation scores for student demographics. They find that this adjustment eliminates the discrepancies in the scores of teachers with different average incoming achievement levels of their students. We run a similar analysis and further test the relationship between unadjusted scores and classroom characteristics in within-teacher comparisons.

We estimate a modified version of Model 2.1 as follows:

$$A_{it} = \alpha + C'_{k(i)t}\delta + \tau_t + \varepsilon_{ikt} \tag{2.3}$$

This Model only controls for classrooms characteristics and year fixed effects and excludes teacher characteristics. We then compute score residuals (or adjusted scores) as

$$A^*_{it} = A_{it} - \hat{A}_{it} = A_{it} - \hat{\alpha} - C'_{k(i)t}\hat{\delta} - \hat{\tau}_t \tag{2.4}$$

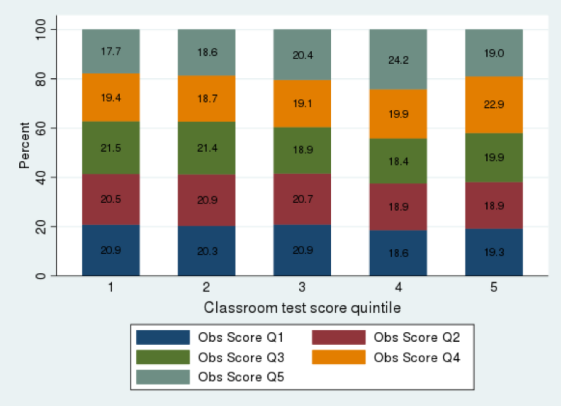
Figure 2.3 recreates Figure 2.2 using the residuals instead of the observation scores.

Panels A and B of Figure 2.3 show that the distribution of adjusted observation scores no longer depends on the classroom's FRL rate and average test scores. For example, 21 percent and 18 percent of teachers in classrooms with the lowest and highest shares of poor students receive a top adjusted score, respectively. The similarity of these percentages means that teachers do not need fewer poor students to have a good chance to receive a high observation score.

Figure 2.3: Quintiles of Adjusted Observation Scores by Quintiles of Classroom Characteristics

(a) Free/Reduced Price Lunch Percentage

(b) Baseline Test Scores



Notes: These figures show the joint distribution of the adjusted classroom observation ratings with FRLP (panel 2.2a) and baseline test scores (panel 2.2b) by quaintiles.

The correlation between adjusted and unadjusted scores is 0.46, which implies that many scores were modified to compensate for differences in classroom characteristics. If we were to lay off teachers at the bottom 5th (10th) percentile based on unadjusted scores, 30 percent (8.6 percent) of them would be misclassified if we used the adjusted scores. Using adjusted scores, however, has some limitations. The adjustment model may fail to account for the sorting of teachers to schools and students. Highly effective teachers could be more likely to be assigned to classrooms with higher achieving and more advantaged students, and this could explain the patterns in Figure 2.2. Adjusting their scores negatively impact them. Another limitation of the model is that it may fail to account for differences in organizational support across schools. As mentioned in the literature review, prior research suggests that

schools in high-poverty areas struggle to train and support their teachers (Kraft and Papay, 2014). This could also explain the patterns in Figure 2.2. The low scores in poorer classrooms may be due to differences in teacher quality.

A second approach to level the playing field is to incorporate student surveys (and possibly from parents and other stakeholders) in the teacher evaluation system. We found that survey scores pass our within-teacher test because they are not correlated with idiosyncratic variation of classroom composition. Looking at the change of the R-squared across specifications in Table 2.6, we find that survey scores are highly reliable. Regressions with only classroom characteristics explain between 0.09 and 0.11 of the variation in observation and survey scores, and regressions with teacher fixed-effects explain between 0.95 and 0.96. Students on average seem to give consistent scores to teachers across years.

2.7 Conclusions

In the spirit of evaluating teachers based on their performance and not for the students they teach, school districts have started to incorporate observation ratings in their evaluation systems. During a classroom observation, the school principal or assistant principal enters a teacher's class to rate his or her performance in multiple areas of classroom instruction and management and provide to him or her feedback about how to improve. In this chapter, we use data from Chicago Public Schools to empirically test the extent to which the students in teachers' classrooms influence the classroom observation scores.

By exploiting within-teacher variation in classroom composition, we provide new evidence on the effects of classroom context on teacher ratings. We find that having higher-achieving students, less disruptive students, and fewer poor students is associated with better observation scores. If year-to-year variation in classroom characteristics is idiosyncratic and orthogonal to unobserved factors that we do not control in regression analyses, these results can be interpreted as causal. At the same time, we do not find that student survey reports of instructional and management experiences with their teachers are sensitive to classroom

context.

Our findings raise questions about fairness of classroom observations, since evaluators seem to penalize teachers with underserved students. One possible approach to solve this issue is to adjust observation ratings by classroom characteristics, as value-added models do, and as we simulate here. This way, the adjusted ratings become orthogonal to these characteristics. However, there are certainly trade-offs to consider. Currently, administrators and teachers see value in classroom observations for providing transparent feedback with clear steps on how to get to the next level of performance. There is beauty in the simplicity of these observations. At the same time, they sometimes criticize regression-adjusted growth measures as a black box with statistical adjustments that are difficult to understand. Policymakers will need to consider how to maintain the perceived usefulness of the observations, while making them fair to all teachers regardless of the student populations and contexts in which they teach.

Another approach may be to include ratings from students (and possibly from other stakeholders such as parents) as additional measures of teacher quality. Our findings in this chapter suggest that student reports of teacher quality remain stable regardless of changes in classroom composition. We note that the student survey reports were not part of the teacher evaluation system, and districts would want to take precautions against any potential manipulation if teachers were held accountable for them.

Ultimately, measuring teacher performance remains challenging, particularly with identifying quality teaching measures independent of the students in the classroom. While much attention has been paid to the sources of bias in value-added or student-growth measures, our findings suggest that a more careful consideration of how teachers are observed and evaluated is also necessary. Perhaps that means more extensive training of evaluators or including the perspective of independent evaluators, or it could mean reconsidering the application of popular classroom observation rubrics across different classroom contexts. Regardless, policymakers should be aware of the potential implications of relying too heavily or solely on a

single measure of teacher practice.

CHAPTER 3

THE MPACT INITIATIVE: USING BEHAVIORAL TOOLS TO INCREASE CHILDREN'S EARLY MATH SKILLS

Co-authors: Susan E. Mayer and Ariel Kalil

3.1 Introduction

Compared to more advantaged parents, disadvantaged parents spend less time and especially less time in educational activities with their children (Guryan et al., 2008; Kalil et al., 2012). Children of low-income parents start school behind children of wealthier parents and the skill gap persists throughout compulsory schooling (Duncan and Magnuson, 2013; Fryer and Levitt, 2006; Heckman, 2006, 2008). Research shows a causal relationship between the amount of time that parents spend in cognitively enriching activities and child outcomes (Berkowitz et al., 2015; Del Boca et al., 2013; Fiorini and Keane, 2014; Hsin and Felfe, 2014). Therefore, understanding the barriers that less educated parents face in spending time promoting their young children's development, and finding ways to overcome these barriers, is both scientifically important and policy relevant.

In this chapter, we focus on present bias as one possible explanation of the gap in parental engagement in educational activities, and math activities in particular, between disadvantaged and more advantaged parents. Present bias is the tendency to overvalue immediate payoffs to a decision (e.g., to spend time in leisure) and undervalue future greater payoffs to a different decision (e.g., to spend time in developmental activities with children), thus leading to procrastination and a failure to invest. Because the payoff to investing in children is realized in the future (e.g. school readiness, high school completion), present-oriented parents may procrastinate and do fewer educational activities with their children. They may instead do activities with immediate payoffs, such as watching television or engaging in other leisure activities with their children. As a result, present-oriented parents underinvest

in their children's development.

Many hypotheses try to explain the difference in investment in children's learning between advantaged and disadvantaged parents. Classic economic theories (Becker and Tomes, 1979) suggest that low-income parents, compared to their wealthier counterparts, may underinvest in their children because they are credit constrained, their children's innate abilities are lower, they have biased or lower expectations of their returns on investment, or they have different preferences for education. At the same time, many math interventions are based on the idea that instructing parents about the importance of math activities and providing information about what parents can do with their children will increase the math interaction between parents and children and hence the math skills of the children (Harackiewicz et al., 2012; Berkowitz et al., 2015; York et al., 2018).

At the Behavioral Insights and Parenting Lab, we designed the Math for Parents and Children Together (MPACT) intervention to test the effectiveness of a low-cost treatment on parental engagement and child's early math skills. This chapter specifically asks (i) whether providing parents with information and math materials increases children's math skills, and (ii) whether a behaviorally informed treatment designed to overcome present bias increases children's math skills beyond providing the information and materials alone.

Present bias has been studied in the domains of finance, labor supply, consumption, and health (DellaVigna, 2009; DellaVigna and Malmendier, 2006; DellaVigna and Paserman, 2005; Jones and Mahajan, 2015; Meier and Sprenger, 2010). However, few studies have investigated the influence of this bias on parental behavior, and even fewer have focused on the behaviors of parents of preschoolers. One exception is the Parents and Children Together (PACT) intervention analyzed by Mayer et al. (2018). PACT is the predecessor of MPACT and was implemented to test the effect of behavioral tools designed to overcome present bias on parental time reading to their preschool children. Parents were provided tablets with a digital library of about 500 books in English and Spanish. Parents were then randomly assigned to a control group or a treatment group that received behavioral tools

(a commitment device, reminders, and a social incentive) in addition to the tablet with the digital library. The study finds that treated parents increased the number of minutes that they read to their children by one standard deviation and the impacts were larger for highly present-oriented parents. A related study that sent text messages to parents of preschoolers is York et al. (2018). They evaluate the effects of the READY4K! intervention, which provided tips and reminders to parents in the form of text messages to help them support their children's literacy development. They find positive impacts on parent's reports of their engagement with their children in literacy activities, of their parental involvement at school, and, in some estimation models, children's literacy skills.

MPACT focuses on early math, a subject that is nearly absent in preschool education, but evidence has shown its predictive power on school success (Claessens and Engel, 2013; Duncan et al., 2007; Watts et al., 2018), choosing STEM careers, or even becoming an inventor (Bell et al., 2018). The mathematical knowledge that children bring to school influences their math learning and, because math is often hierarchically organized, a weak foundational math knowledge hinders the acquisition of more advanced math skills (Baroody et al., 2012; Purpura and Lonigan, 2015). Individual differences in math skills emerge early; hence, the goal of MPACT is to improve the early math skills of low-income children.

MPACT is a 12-week intervention done in three rounds, each with about a third of the sample. Participant parents and their children were randomly assigned to a control group, MKit group who received information and materials in the form of a math activity booklet, or MKit + present bias group who received behaviorally informed text messages in addition to the booklet. Children's math skills were assessed at the end of the 12-week intervention and at 6 and 12 months after the intervention ended. We currently have data on the full sample on the math assessment at the end of the intervention, but do not have data on the follow-up assessments for all rounds. Thus, we analyze the data on math scores immediately following the 12-week intervention for more than 1,400 children from low-income families enrolled in 29 Head Start centers in Chicago.

At the end of the intervention, parents report more math engagement with their children and greater investment in math-related materials. However, the analyses show no statistically significant (at $p = 0.05$) improvement in math skills for either the MKit or the MKit + present bias groups immediately after the intervention. We also do not find changes in parental math attitudes and beliefs. We test several hypotheses for the null short-term treatment effects, including the influence of peer effects. Subgroup analyses show that Spanish-speaking families benefit from the intervention.

This chapter makes three main contributions to the behavioral economics and economics of education literatures. First, it is one of the few math interventions that employs behavioral tools in the home environments of low-income families. Second, we account for peer effects with our two-stage randomization protocol. Third, we complement the existing set of time-preference studies by providing estimates of the time-preference parameters of low-income parents of preschoolers and reporting heterogeneity within this group.

This document is organized as follows. Section 3.2 describes the research design of MPACT, including the randomization protocol, treatment arms, and description of data. We present descriptive statistics and balance test in Section 3.3. The next section develops the empirical strategy and presents main results of treatment effects. Section 3.5 discusses the results. We conclude in the final section.

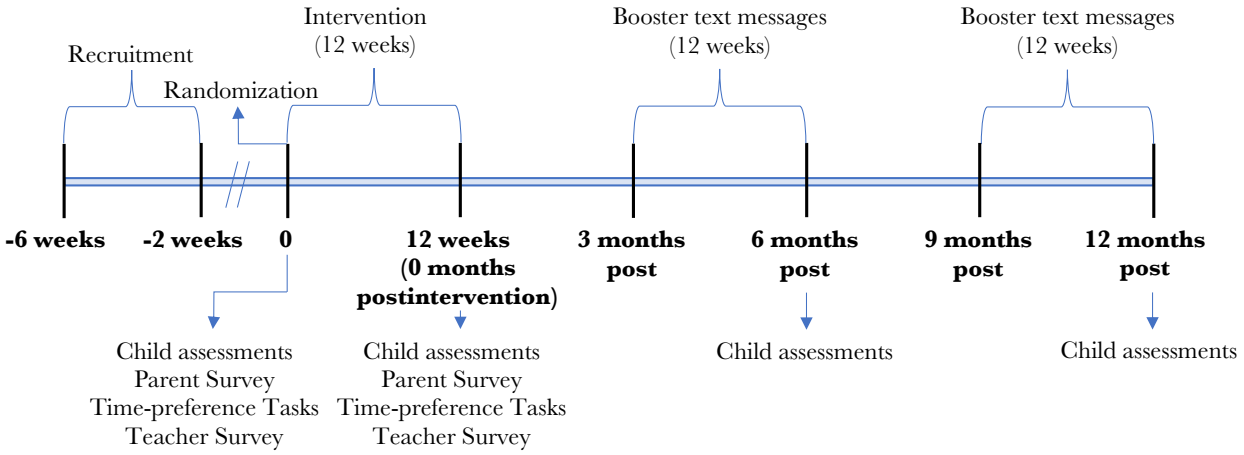
3.2 Research Design

3.2.1 Recruitment

M-PACT is a 12-week intervention aimed at improving parental engagement in math activities with their children and improving children's math skills. It was conducted in three rounds: round one started in fall 2017, round two started in spring 2018, and round three started in fall 2018. Each round followed the timeline in Figure 3.1. About four weeks were spent recruiting families, after which random assignment to treatment groups was conducted

and the baseline child assessments and parent and teacher surveys were carried out. The intervention lasted for 12 weeks, then the first follow-up assessments and surveys were implemented. 6 and 12 months after the intervention, we collect the longer-term child assessments, which is still ongoing.

Figure 3.1: Timeline of the MPACT Initiative



We recruited 29 preschool centers that serve 3 to 4-year-old children attending Head Start programs that are either community based or operated by the Chicago Public Schools. These centers are located primarily in the south, west, and northwest areas of Chicago in lower income neighborhoods. These centers were selected for several reasons: they represented the diversity of centers in Chicago, they are medium or large in size, and they are located in areas accessible for our research assistants. We made efforts to select centers with racially and ethnically diverse populations, so that our sample of children mirrors the population of Head Start children in Chicago. We recruited 1,458 children and 1,410 parents in these centers.

We recruited parents in one of two ways depending on the preschool.[1] Our preferred

1. Only the principal guardian was chosen to participate. We had multiple ways of identifying the principal guardian. Usually we asked the centers which parents picked up or dropped off the child most often, that way we could ensure they could receive materials and take the surveys. We also called all the parents who were duly enrolled to let them know that we chose the "other parent" as the primary parent for MPACT. In some other cases in which the parent who initially enrolled the child was not the one who spent most of the time with the child or came to school regularly for picking up and dropping off, we offered the other parent to sign the consent "to transfer the duty" ("making void" the initial consent). At the same time, we made

method for recruitment was an opt-out model, in which all eligible parents were automatically enrolled in the project but were provided with an array of opportunities to opt-out of participation. In centers that did not approve an opt-out model, we adopted an opt-in model, in which research assistants recruited parents in person by approaching them when they dropped off and picked up their children. Twelve preschools allowed opt-out recruitment. Participation rates were high at both opt-out (99 percent) and opt-in (71 percent) preschool centers.[2] Of 1,458 participant children, 43 percent were in opt-out preschools.

All parents received either a consent or opt-out form (depending on the recruitment type) approved by the University of Chicago Institutional Review Board and the Chicago Public Schools Research Review Board. To be eligible to participate in MPACT, the child had to be between 3 and 4 years old, the parent had to speak English or Spanish, and the parent had to have a working cellphone and be willing to receive text messages. Younger children were not included primarily because the math skills that we measure are not reliably assessed for younger children. After a parent consented to participate, he or she and the child were randomly assigned to the control group or one of the four treatment arms (explained below). The MPACT field experiment is registered in the American Economic Association Registry for Randomized Controlled Trials and the unique identifying number is AEARCTR-0002512.

3.2.2 Treatment arms

All treatments were offered in Spanish and English, depending on language of choice for each family.

Control group. Parents received a placebo treatment, which was a reading book. They also received administrative text messages, such as welcome and thank you messages and reminders to complete the surveys.

---

clear that only one parent was going to receive the text messages and complete the surveys. That did not happen a lot, but parents appreciated this flexibility.

2. The number of eligible parents was provided by preschool centers; it is the number of parents who met the eligibility criteria for opt-out centers or the number of children in the classrooms for opt-in centers.

MKit only group. Parents received information and materials in the form of a math activity booklet. We designed the booklet, which we named MKit, with 22 developmentally appropriate math activities for parents and children to do together. It also included a game board, a number grid, game pieces, and a goal tracker. The math activities focused on five specific skill areas within the numeracy domain: number recognition, counting, comparing size and quantity, adding and subtracting, and patterns.

Each activity card provided instructions for how to do the activity and included tips for how to get the most out of the activity—for example, encouraging parents to use math words like *more than* and *less than*. The activity cards also provided suggestions for how to make the activity harder after the child mastered the initial activity. One example activity from the MKit is shown in Figure 3.2. The activity is to ask the child to count as high as he or she can and integrate a physical activity while counting. To make it harder, the child is asked to count backward. The Math is Everywhere activity—a tip to see math in everyday activities—is to clap and count the squares of the sidewalk or the stairs in the home.

The booklet had instructions on how to use it and some information on the importance of parents spending time in math activities with their child for the child's future success. The content of the MKit was extensively piloted to assure that parents understood the activities and were enthusiastic about doing them.

This treatment arm tests whether simply providing information and materials to parents can increase parental engagement and thus child's math scores. This could happen because the materials decrease the costs to parents of engaging in math activities because the materials are readily available and require little effort on the part of parents to think of or organize activities. The material could increase the efficiency of parental time in math activities as parents learn what activities increase child's math skills. Or it could also increase understanding of the importance of math activities to future outcomes, hence raising expected returns to engagement.

MKit + present bias group. Parents received an MKit and about four behaviorally in-

Figure 3.2: Example of a Math Activity in the MKit

Clap a bunch

What your child will learn:
Counting out loud

STEP 1

Ask your child: "How high can you count? Show me with claps! Start at 1 and clap each time you say the next number." Have your child clap once for each number. If your child misses a number, correct her and let her keep going.

STEP 2

Repeat this game with jumps, hops, stomps, or another movement.

Make it Harder

Once your child can easily count to 20, have your child count backward. Start by going backward from 5, then backward from 10 and so on, clapping for each number. Or have your child count by 2s, 5s or 10s clapping on each number. For example, have your child clap on 2, 4, 6, 8, and so on.

Math is Everywhere

The world is full of things to count. Clap and count the squares on the sidewalk, the stairs in your home, or the steps it takes to get to your child's bed or another place in your home. Putting physical activities together with counting helps children learn numbers.

formed text messages per week that were intended to overcome the procrastination associated with present bias. Although the wording of the messages differed each week, the primary content did not. One text message per week directed parents to set a goal for how many days they would do math activities with their child and encouraged them to write and track their goals. The messages did not direct parents to engage in specific activities. Nor did they suggest what the goal should be or how often the activity should occur. Setting a goal works as a soft commitment device to overcome present bias by increasing the psychological costs of not sticking to the goals (Bryan et al., 2010). Another text message suggested that parents share their goals with a friend or relative, which could also work as a soft commitment device by increasing the social cost of not achieving the goals. Another text message "brought the future to the present" by highlighting the importance to the child's future of engaging with the child now. The rationale of this message was to make future payoffs more salient in the present, thus mitigating the undervaluation of these payoffs. Finally, one or two text messages per week were sent to remind parents to work toward their goal. These texts addressed inattention and forgetfulness.

The MKit + present bias treatment arm was intended to increase parental time on math activities, and consequently improve children's math test scores by overcoming present bias. It was not necessarily intended to change parents' discount on the future but it was intended to help parents adapt to the procrastination that arises from discounting the future. The MKit + present bias, MKit, and control groups are the focus of this chapter.

Other treatment groups. The other treatment arms were *MKit + growth mindset group* and *tablet group*.[3]

3. Parents in the MKit + growth mindset group were provided an MKit and received text messages with growth mindset content. Parents in the tablet group were provided a tablet with educational math apps. In addition, before the 6- and 12-month follow-up assessments, parents in the MKit + present bias and MKit + growth mindset arms received booster text messages for 12 weeks. Analyses of these treatment arms and booster text messages will be presented in a different paper.

3.2.3 Randomization and power analysis

We randomized in two stages. In the first stage, we randomly assigned classrooms across schools to either a treated or untreated group. We assigned 15 classrooms (5 in each round) to the untreated group. All students in these classrooms were assigned to the control group. Next, we used block randomization in the treated classrooms, randomly assigning students to the control, MKit, MKit + present bias, or to the other treatment arms.[4] We implemented a two-stage randomization to study peer effects.

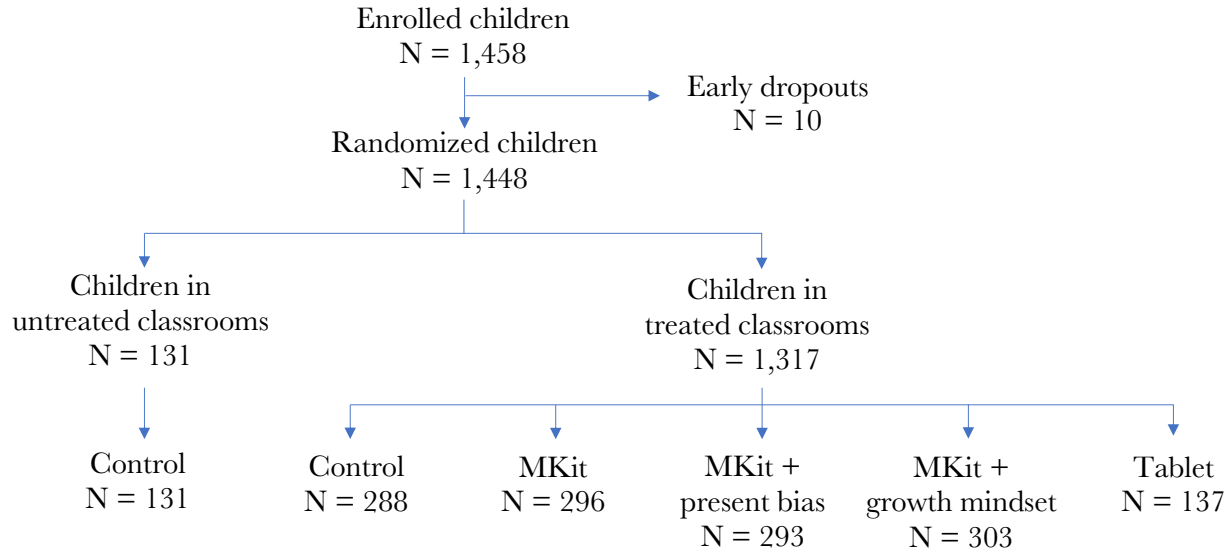
If a parent had more than one qualifying child in the preschool, all the siblings were put in the same treatment group. If any sibling was in an untreated classroom, then all the siblings were assigned to the control group. Otherwise, we used simple randomization to assign sibling groups to control, MKit, MKit + present bias, or growth mindset arms. Siblings were not placed in the digital tablet group because only one tablet was provided per family, and this could have caused conflict and obvious spillovers between siblings. The number of sibling groups who enrolled in MPACT was 47, for a total of 95 children. The analyses reported here include siblings, but excluding them does not change results qualitatively.

The number of enrolled children by treatment group is shown in Figure 3.3. While 1,458 children were enrolled, 10 children dropped out before randomization leaving 1,448 children who were randomized, 133 of whom were assigned to untreated classrooms and 1,315 to treated classrooms. All children assigned in untreated classrooms are assigned to the control group. Among children in treated classrooms, 286 were assigned to the control group, 296 to the MKit group, 293 to the MKit + present bias group, and the others were assigned to other treatment arms not discussed here.

Power analysis estimates indicated that treated children in the MKit and MKit + present bias groups needed to score 0.20 standard deviations higher than control group children for our estimates to detect a statistically reliable difference. These estimates were obtained

4. Each group had a similar sample size within classrooms with the exception of the digital tablet group, which was about half the size of the other groups.

Figure 3.3: Number of Participant Children by Treatment Group



Notes: Enrolled children are those whose parents consented to participate in opt-in centers or did not express their unwillingness to participate in opt-out centers.

through Monte Carlo based simulations following the two-step randomization protocol and assuming a sample size of 1,500 children in 150 classrooms (10 children per class), intraclass correlation of 0.2, power of 80 percent, and significance level of 5 percent. A 0.2 effect size is standard in the education literature, and it is equivalent to moving a 3- or 4-year-old child from the 42nd percentile rank to the 50th in the Woodcock-Johnson IV Applied Problems, the main test we use in MPACT.

3.2.4 Data

The main outcome for MPACT is the child's numeracy skill, as measured by the Woodcock Johnson IV Applied Problems (hereafter WJ). The WJ subtest focuses on elements of numeracy, such as counting, number recognition, and addition and subtraction. MPACT was tailored to improve numeracy skills.[5] The assessments were conducted in English, Spanish,

5. Other math competencies that we did not measure are geometry, measurement and data analysis, spatial competencies, patterning, among others (Nguyen et al., 2016).

or both languages depending on each child's needs. Bilingual assessors with experience in assessing children were hired and trained to do the assessments.

We also conducted parent surveys at baseline and at the end of the intervention.[6] Parents responded to questions about their child's characteristics (health, Individualized Education Program), parental demographics (sex, age, race and ethnicity, education level, employment status, enrollment at a school), household characteristics (family income, family composition), attitudes toward math (math anxiety, enjoyment doing math, knowledge on helping child to get ready for kindergarten math), time investment on child, and beliefs and expectations about the math skill production function. In addition to surveys, parents completed tasks designed to measure parents' time discount factors. We explain these tasks below, but first we define our primary outcome and measures of mechanisms.

WJ z-score

The primary outcome is the WJ z-score, which is the raw score adjusted by age and norm referenced with respect to a representative sample of the U.S. A score of zero indicates that a child scored as the average child for his or her age. The mean score for MPACT children, however, is -1.1, which indicates that the mean score for children in our sample scored was more than one standard deviation lower than the average child in the U.S. To put it another way, our participant children ranked in the 23rd percentile of the test score distribution. This shows that a gap exists even before children start formal schooling, a finding consistently documented in the literature when comparing disadvantaged children with their more advantaged counterparts (Fryer and Levitt, 2006, 2013; Magnuson and Duncan, 2016; Reardon and Portilla, 2016).

6. We also collected information on classrooms through teacher surveys, but we do not analyze these data in this chapter.

Math home environment

In the parent surveys, we asked a series of questions about the child's home environment and parental attitudes and beliefs about math in order to construct measures of mechanisms that we believe might account for changes in test scores and that might be altered by the intervention. Our hypothesis is that the treatment will improve these measures and lead to higher parental engagement, which in turn will raise the child's math scores.

We have five outcome measures: an index of barriers to do math activities, an index of math activities in the home, an index of reading activities, an index of material investment, and an index of math time in the home. The first three indexes are measures in both the baseline and follow-up periods, while the latter two are measured in the follow-up period only. For indexes constructed at both baseline and follow-up, we employed principal component analysis on responses to the *baseline* survey in order to extract the factor loadings of the first principal component. Then, we used these factor loadings to construct the indices at both baseline and follow-up.[7] For the indexes available only in the follow-up, we used the factor loadings of the *follow-up* survey. All indexes are normalized to have zero mean and unit standard deviation.

Below we briefly describe each index.

Index of negative barriers to do math activities. This index captures barriers that impede parents to engage with their children. Parents were asked in a 6-point Likert scale the extent to which they agree or disagree with the following seven statements: (i) I have all the time I need to help my child to get ready for kindergarten math, (ii) I am not sure what my child needs to learn to get ready for kindergarten math, (iii) Sometimes I am too tired to play with my child in ways that would help her learn math, (iv) I have everything in my home that I need to help my child get ready for kindergarten math, (v) Math is hard for my child, (vi) I don't like math, and (vii) I avoid doing math activities with my child because she does

7. The treatment may have affected the relationship across the variables that compose each index, hence changing the factor loadings. By fixing the loadings at baseline, we study changes in the amount reported for each variable.

not pay attention or sit still for long. Responses were between 1 (strongly disagree) to 6 (strongly agree). The responses for statements (i) and (iv) were reverse coded so that a high value had a negative connotation. Correlation of index of negative barriers and baseline WJ is -0.14 ($p < 0.01$).

Index of math activities. We asked parents how often over the last week they helped their children count, recognize numbers, read books that have numbers, add or subtract, and recognize shapes. Additionally, we asked parents how often their children used educational apps or games on a digital device and how often they watched educational television. Responses were "never" (with a value of 0), "1 or 2 days" (value of 1), "3 or 4 days" (value of 2), and "5 or more days" (value of 3). This index is positively correlated with baseline WJ ($\rho = 0.10$, $p < 0.01$).

Index of reading activities. As in the index math activities, we asked parents to indicate how often they helped their children learn the sounds of letters or learn to read and recognize letters. We use this index to see if the math treatment had spillovers to parents' reports of reading time. This index and baseline WJ are positively correlated ($\rho = 0.10$, $p < 0.01$).

Index of material investment. Parents were given a list of items that children could use to learn math, and they indicated whether they had owned them. These items included toy blocks or Legos; children books with numbers; child board games or card games that have numbers or use dice; toys for learning math (e.g., abacus, dominoes, etc.); apps, CDs, or DVDs for learning math; and a computer or electronic tablet that child can use. Another option, "other things not listed here," was in the list but excluded from the index. This index is constructed only at the first follow-up.

Index of math time. In the follow-up survey, we asked parents how often they helped their children do math activities and learn math in everyday activities since MPACT started. Responses were "never" (value of 0), "1 time a month or less" (value of 1), "2 or 3 times a month" (value of 2), "1 or 2 times a week" (value of 3), "3 or 4 times a week" (value of 4), and "5 or more times a week" (value of 4). Because these questions were asked only at

the *end* of the intervention, this index is separate from the index of math activities, which was measured at both baseline and follow-up. Therefore, it is a complementary measure of parental engagement in math time.

Math anxiety, growth mindset, and parent efficacy

Measures of parental attitudes and beliefs hypothesized to influence children's test scores were measured at baseline and follow-up. Parent's math anxiety has been shown to have a negative relation with children's math achievement (Maloney et al., 2015). We find a small negative correlation in our data ($\rho = -0.07$, $p = 0.05$). To measure math anxiety we asked parents to rate how nervous math makes them and give a number between 0 (not at all nervous) and 10 (extremely nervous).

Another factor hypothesized to influence a child's learning is whether the parent has a growth mindset. While a parent with growth mindset believes that intelligence is malleable, a parent with a fixed mindset believes people can learn new things but their underlying intelligence remains the same. Consequently, parents with a fixed mindset would invest less in their children than parents with a growth mindset because their investments would have little impact on child's intelligence, according to their beliefs. We employ the Dweck scale to measure growth mindset. This scale has eight items, and in each item respondents specified how much they agreed on a scale of 1 (strongly disagree) to 6 (strongly agree) to statements such as "people have a certain amount of intelligence, and they can't really do much to change it" and "to be honest, a person can't really change how intelligent they are." Responses were recoded so that a higher score represents higher growth mindset and a lower score represents a fixed mindset. Studies have shown that students with a growth mindset have higher achievement than their peers (Blackwell et al., 2007; Claro et al., 2016). In our data, we find that a parent's Dweck score is positively associated with the child's baseline math test score ($\rho = 0.12$) and this relationship is statistically different from zero ($p < 0.01$).

A different yet related measure to growth mindset is parents' expected return to the time

they spend in math activities with their children. Parents who expect a higher return are likely to spend more time in math activities compared to parents who expect a low return. We asked parents how they thought their child's math skills in kindergarten would compare to the math skills of children in kindergarten across the U.S., using a scale from 0 (far behind) to 10 (far ahead). This is the reference question, and it has a statistically significant positive correlation with baseline WJ ($\rho = 0.20$) and Dweck score is ($\rho = 0.16$). Then, in a follow-up question, we asked parents what they thought their child's rank would be if they were to spend 15, 30, 45, or 60 minutes more every week doing math activities with their child. We compute the marginal effect as the change in child's ranking from the reference question to the question about 15 more minutes. We standardize this measure to have mean zero and standard deviation of one to ease interpretation.

Present bias measures

Our main hypothesis about why some parents spend less educational time with their children is that they are prone to impatience or that they have a high discount rate for the future. To identify parents who are impatient, we administered a task to estimate the time-preference parameters for money.

The Money Game is an adaptation of the Convex Time Budget first introduced by Andreoni and Sprenger (2012). Parents were asked to choose between an amount of money that they could receive *soon* or a different amount that they could receive at a *later* date. Specifically, participants answered 15 questions with four choices each. In the first set of five questions, parents made inter-temporal choices between receiving payments *now* or *3 weeks* later, the next set of five questions gave parents the choice of receiving payments *now* or *6 weeks* later, and the last set of questions offered the choice of receiving payments in *3* or *6 weeks*. Within each set, each question presented an increasing price for the earlier payment.

To induce participation, respondents were told they would enter a lottery and have a chance to win up to $24. To induce truthful responses, one of their answers was randomly

selected and implemented. Payments were based on parent's responses; thus, for example, if the response to the randomly selected question was $10 in 3 weeks and $8 in 6 weeks, then we would give the winner the respective amount of money on those weeks.[8]

Appendix C.1 explains our estimation strategy. In short, we assume that individuals have a quasi-hyperbolic utility functions, also known as $\beta\delta$ or time-inconsistent utility.[9] Variation in the timing *now* versus *later* helps identify present bias (or β parameter) and variation in the delay or interval between payments—that is, *3 weeks* and *6 weeks*—helps identify the discount factor (or δ parameter). A parent is classified as impatient or highly discounting the future if her estimated discount factors, $\beta \times \delta$, is below the median.

3.3 Sample Selection, Descriptive Statistics, and Balance Test

3.3.1 Sample selection and child assessment and survey completion rates

We randomized 1,448 children. After randomization a center with 25 children closed. We were able to include 4 of these children because they switched to another center that was participating in the RCT. Another 13 parents dropped out after randomization; they did not participate in any parental surveys and their children were not assessed at any point in time. This left 1,414 children, which is our main sample.

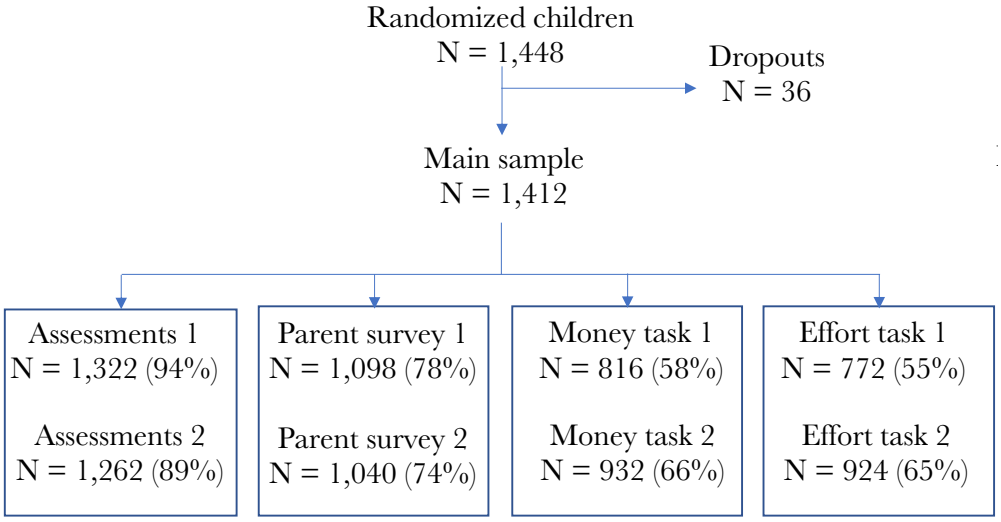
Completion rates of the assessments, parent surveys, and time-preference tasks for the main sample are shown in Figure 3.4. Of 1,414 children, 94 percent and 89 percent were assessed at baseline and follow-up, respectively. 85 percent of these children have both baseline and follow-up assessments. 78 percent of parents completed the baseline survey and

8. In addition to the money task, a time-preference effort task was also administered, and although they were not strongly correlated, we found no major differences in results when using the effort task. This effort task was a modified version of the instruments used by Carvalho et al. (2016) and Meier and Sprenger (2015), and it asked participants to make 12 intertemporal choices on a multiple time list. Parents were asked to choose whether they preferred to spend X minutes *soon* or 30 minutes at a *later* date answering an opinion survey, where X took values 10, 15, 20, and 25. The first set of 4 questions was between *now* and *3 weeks*, the second set of 4 questions was between *now* and *6 weeks*.

9. Specifically, the discount factors of the utility is $1, \beta\delta, \beta\delta^2, , \beta\delta^3, \ldots$ for $t = 0, 1, 2, 3, \ldots$, where β and δ are between zero and one. The parameter β captures present bias and δ is the standard discount factor.

74 percent the follow-up survey. 58 percent of parents took the baseline time-preference task and about 66 percent took the follow-up task. The increase in the take-up of the money task is because we made this task "not optional" in the follow-up survey. In other words, at baseline, parents were asked if they were willing to participate in these tasks for a chance to win a monetary reward, while at follow-up we omitted this filter question so that the tasks were shown to everyone taking the survey. Siblings share the same parental responses.

Figure 3.4: Child Assessment and Parent Survey Completion Rates



Notes: Assessments 1, parent survey 1, money task 1, and effort task 1 correspond to baseline. Those with number 2 correspond to the follow-up period.

We test for differential attrition, assessment completion, and survey response, and present results in Table 3.1. The tests come from separate regressions where the dependent variables are indicators for whether a child dropped out, completed the assessments, or whether their parents responded to the surveys or the time-preference task. The explanatory variables are treatment group indicators and classroom fixed effects, given that the randomization was done within classrooms. Looking at individual treatment coefficients (columns 3 and 5), no coefficient is statistically significant at 5 percent significance level and only one coefficient is marginally significant at 10 percent. Column 7 reports results of a joint significance test, in which the null hypothesis is that the MKit and MKit + present bias coefficients are equal to

zero. We cannot reject this null hypothesis. Taken together, these results show no evidence of differential attrition or completion rates across any of our treatment comparisons.

Table 3.1: Differential Attrition, Assessment Completion and Survey Response across Treatment Groups

	Control		MKit		MKit + present bias		Joint p-value	N
	Mean	S.D.						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dropped out	0.02	[0.15]	0.01	(0.01)	-0.00	(0.01)	0.49	1448
Completed assessment 1	0.94	[0.24]	0.01	(0.02)	-0.01	(0.02)	0.58	1412
Completed assessment 2	0.89	[0.32]	0.03	(0.03)	0.05*	(0.03)	0.15	1412
Took survey 1	0.76	[0.43]	-0.02	(0.04)	0.03	(0.04)	0.30	1412
Took survey 2	0.72	[0.45]	-0.01	(0.04)	0.00	(0.04)	0.96	1412
Took money task 1	0.57	[0.50]	-0.03	(0.04)	0.01	(0.04)	0.63	1412
Took money task 2	0.64	[0.48]	-0.01	(0.04)	-0.03	(0.04)	0.76	1412
Invalid assessment 1	0.04	[0.21]	-0.01	(0.02)	0.00	(0.02)	0.77	1412
Invalid assessment 2	0.03	[0.17]	-0.01	(0.01)	0.02	(0.01)	0.09	1412

Notes: Each row is a separate regression where the dependent variable is indicated in the first column and explanatory variables are treatment group indicators and classroom fixed effects. Each dependent variable takes value one if the individual dropped out, completed the assessments, responded the surveys or took the time-preference task. The first regression has $N = 1,448$ (all children who were randomized) and the following regressions exclude dropouts and have $N = 1,412$. Columns 1 and 2 show the control mean and standard deviation (in brackets). Columns 3 and 5 show estimated treatment coefficients and columns 4 and 6 report robust standard errors (in parentheses). Column 7 reports the p-value from a joint significant test of the MKit and MKit + present bias coefficients. Column 8 reports the number of observations. Asterisks denote significance level: * $p < .10$, ** $p < .05$, *** $p < .01$.

For the rest of the analyses, we imputed missing values for test scores and survey items of children in our main sample. For children who did not take the baseline or follow-up assessments, we predicted their raw test scores based on their sex, age, and language. That is, we assign the same adjusted score based on these characteristics irrespective of treatment group.[10] We created indicators for these imputed cases to include in regression analyses. As a robustness check we exclude imputed test scores from analyses. In regard to surveys, if a survey taker did not answer a specific question or item, we set it to zero if the variable is dichotomous or to the median value of other respondents in the same treatment group if

10. 6 percent of children in our main sample had no baseline scores and 11 percent had no follow-up assessments. 63 children had missing baseline assessments only, 123 children had missing follow-up assessments only, and 29 children had both assessments missing.

the variable is continuous.[11] We include an indicator for missing values. We did not impute responses for parents who did not take the survey.

Assessors identified children who took the assessment but who had communication difficulties or cognitive disabilities for whom a valid assessment score could not be given.[12] Some test scores were identified as invalid: 4.4 percent of baseline test scores and 2.8 percent of follow-up test scores. There is no evidence of differential invalidity across treatment groups, as shown in the last two rows of Table 3.1. The joint significance test when baseline invalid score is the dependent variable yields a p-value of 0.77. Although the p-value for follow-up invalid scores is marginally significant ($p = 0.09$), the individual treatment coefficients are undistinguishable from zero. Treating invalid scores as valid may generate measurement error because children with disabilities may have guessed their responses. Therefore, we set the invalid raw scores to the mean and create an indicator for these cases.[13]

The sample size will vary depending on the model specification. For our main analysis, the sample size is 1,412, who are children with baseline and follow-up assessments. In analyses that use baseline parental and family characteristics, sample size is 1,098. Analyses that use information from the end-line survey have a sample size of 1,040. Finally, specifications that utilize responses to the time-preference task have fewer cases. Given that we do not find evidence of differential attrition and survey take-up across treatment groups, these subsample analyses are internally valid.

11. Results are qualitatively unchanged whether we drop the missing cases or impute the median value of all responses irrespective of group assignment.

12. Children with communication difficulties and cognitive disabilities were identified based on their test scores and assessors' notes. First, children with WJ scores of 3 or lower (i.e., very low proficiency scores) were selected. Then, the assessors' notes and math test scores were examined for indication of communication difficulties or cognitive disabilities that suggest the assessment score may not be valid. Two RA reviewed invalid scores and a third person resolved any discrepancies. They were asked to discern between communication/cognitive difficulties and math competency issues (e.g., a child not answer questions because they have obvious cognition issues versus not knowing enough math to understand what the BTA is asking).

13. Results are qualitatively unchanged whether we drop these cases or keep the scores they initially had

3.3.2 Descriptive statistics and representativeness of main sample

Descriptive statistics of our sample and Head Start children in the U.S. and Chicago are reported in Table 3.2. Columns 1 and 2 present the mean and standard deviation for our sample. Column 3 shows data from the Head Start Family and Child Experiences Survey (FACES) 2014. Column 4 comes from administrative data of the Office of Head Start at the U.S. Department of Health and Human Services. Column 5 also comes from the Office of Head Start but is restricted to Chicago. We employ FACES 2014 because it has more information of children and their families than the administrative data; however, it cannot be disaggregated by city.

In comparison to a national sample of Head Start children, our sample is younger primarily because we confined our sample to three-year-olds except in CPS schools where we extended the age to four. In terms of ethnic and racial composition, we have more Hispanic and fewer white children than a national sample (columns 3 and 4) and more Hispanic and fewer black children than Chicago Head Start children (column 5). However, there are no large differences in the language spoken at home across samples.

The mean of the WJ standard score in our sample is 83.8, which is lower than the U.S. average of 93.1. The WJ standard score is norm-referenced to have a population mean of 100. The standard score mirrors the z-score, with the difference that the latter is expressed in standard deviations.[14] Both WJ standard score and z-score shows that our sample has lower math skills than the average U.S. child. Finally, about 10.8 percent of our sample has an individualized education program or disability and 2.4 percent had low birth weight.

Parents of MPACT children differ from Head Start families in some characteristics. Most participant guardians are female. They are also younger, slightly less educated, and less likely to work full time than national averages. They live with the child's biological parent at similar rates as the U.S. average. The percentage of families living in poverty is higher

14. The between- and within-classroom standard deviations of WJ z-score are 0.51 and 0.89, respectively. This means that 25 percent of the total variation in scores is due to differences across classrooms while 75 percent is due to differences within classrooms.

than the U.S. but lower than Chicago.

3.3.3 Balance test

We perform balance tests of treatment groups on several baseline characteristics, some of which come from the consent forms and child assessments ($N = 1,414$) and others from the baseline parent surveys ($N = 1,098$). Appendix Table C.2.1 presents the results of these tests and we present their graphical representation in Figure 3.5 below. For each baseline characteristic, we run a separate regression with treatment indicators and classroom fixed effects and test for joint significance. Each dot in the graph denote the coefficient point estimate and the horizontal lines represent the 95 percent confidence interval. Most of the variables reported in the graph are balanced across treatment groups, with the exception of the WJ z-score. MKit participants scored 0.15σ higher than the control group and this difference is significant ($p<0.10$). Due to this unbalance, the joint-significance test is rejected at conventional levels. We thus control for the baseline scores in our regression analyses.

Other baseline characteristics—including whether a parent's highest education is high school, whether a parent has some college education, and whether a child had low birthweight—are individually marginally significant ($p<0.10$) for either treatment group, but we cannot reject the joint significant tests. Additional variables are reported in Appendix Table C.2.1. Out of 28 comparisons, the joint significance test is rejected for two cases, which is about one would expect if 5 percent of cases were unbalanced by chance. It is worth highlighting that all these variables, with the exception of child's sex, age, and language of the consent form, were collected after the randomization occurred. To deal with potential imbalances, we control for parent and family characteristics.

Table 3.2: Descriptive Statistics of Main Sample and Head Start Children in the U.S. and Chicago

	Main sample		Head Start FACES 2014	Head Start U.S. 2017	Head Start Chicago 2017
	%	S.D.	%	%	%
	(1)	(2)	(3)	(4)	(5)
Child characteristics					
Male	48.1		49.6		
3 years old or younger	57.3		44.5	47.2	47.9
4 years old or older	42.7		55.5	52.8	52.1
White non-Hispanic	2.5		27.7	24.2	2.8
Black non-Hispanic	31.1		22.2	30.2	51.2
Hispanic	63.3		41.8	36.3	41.5
English primary home language	74.8		76	72.1	75.4
Spanish primary home language	24.3		21.4	22.1	21.7
WJ standard score (mean)[a]	83.9	[15]	93.1		
WJ z-score (mean)	-1.1	[1]			
Individualized Education Program	10.6				
Low birthweight (less than 3.5 pounds)	2.4				
N (children)	1,414		1,908	798,306	19,484
Parent and family characteristics					
Female principal guardian	92.3				
Age (mean)	31.2	[6.9]			
18-29 years old	44.3		12.8		
30-39 years old	43.5		29.6		
40-49 years old	10.7		23.1		
50-59 years old	1.3		23.4		
60 years old or older	0.3		11.1		
Less than HS diploma	24.8		22.4		
HS diploma or GED	28.3		34.2		
Some college	34.2		34		
Bachelor's degree or higher	7.2		9.5		
No. of people in household (mean)	5.3	[2.1]	4.4		
No. of children (5 years old and younger)	1.6	[0.8]			
No. of adults (18 years old and older)	2.2	[0.9]			
Live with child's biological parent	46.4		46.9		
Working full-time	21.7		52.1		
Working part-time	31.0		22.4		
In school	15.6				
Below 100% of federal poverty line[b]	73.7		67.7	71.2	78.5

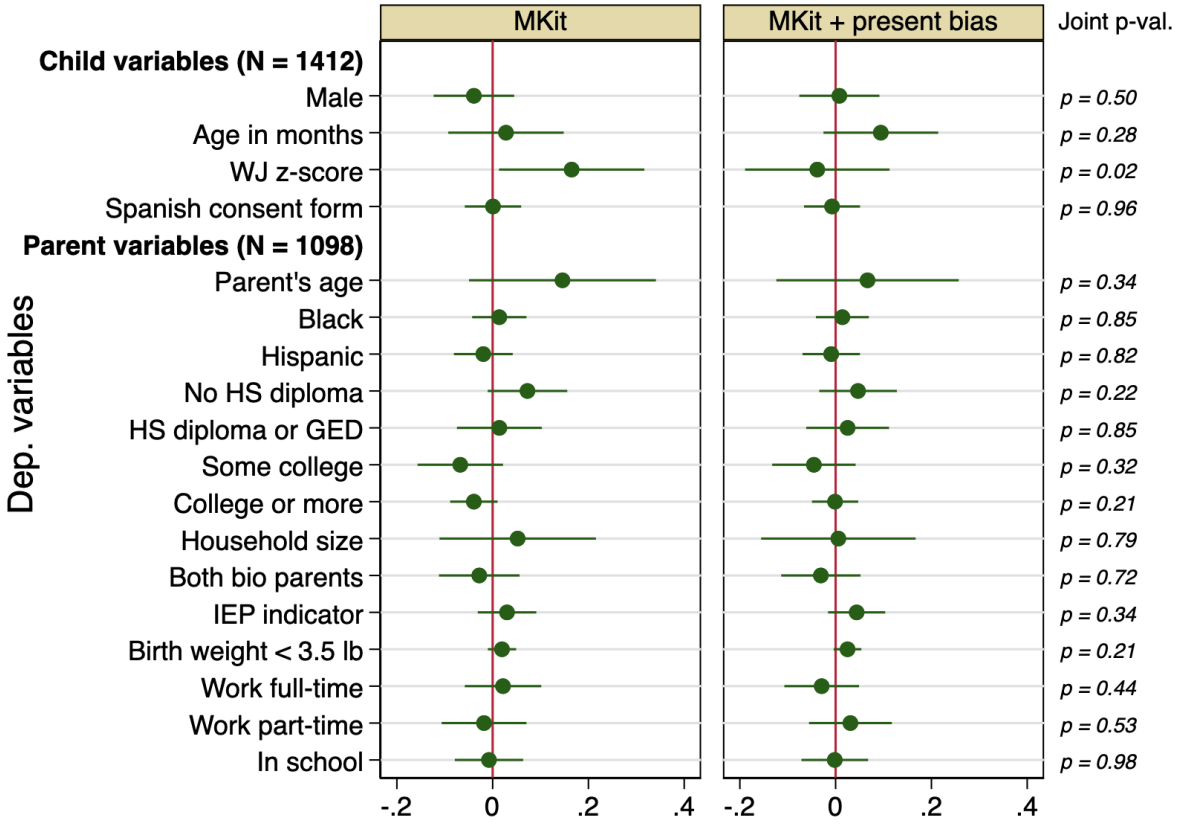
Sources: FACES 2014 data come from Aikens et al. (2017) and Moiduddin et al. (2017). Head Start U.S. and Head Start Chicago data come from administrative data published in the Office of Head Start Performance Indicator Report (2017).

Notes: Proportions are reported except when otherwise indicated. Standard deviations are in brackets. Main sample's race, ethnicity, language, IEP status, and birthweight come from the parent surveys. The percentages of parental education in the main sample do not sum to one because of missing/don't know responses.

[a] Standard score is the WJ-IV Applied Problems for the main sample and WJ-III Normative Update Applied Problems for FACES 2014.

[b] Estimates for the main sample are based on reported family income, family size, and federal poverty line.

Figure 3.5: Balance Test on Baseline Characteristics



Notes: This graph shows balance tests of baseline characteristics. Each row comes from a separate regression, in which the dependent variable is the indicated variable on the first column and the explanatory variables are treatment group indicators (with the control group as the omitted category) and classroom fixed effects (given that randomization was done within classroom). The first set of regressions has N=1,414 and the second set has N=1,098 because they use responses from the parent survey. The dots in the graph are point estimates and the horizontal lines depict their 95 percent confidence intervals. P-values from joint significance tests that the MKit and MKit + present bias coefficients are zero are reported on the last column. The variables child's age, parent's age, and household size were normalized to fit in the graph, but the reported p-values are from the untransformed data.

3.4 Short-Term Treatment Effects

This section presents the empirical strategy to estimate the treatment effects of the MKit and MKit + present bias interventions. It also shows results of the impacts on the home environment outcomes, children's early math skills, and heterogeneous effects by variables preregistered in our analysis plan.

3.4.1 Empirical model

Let Y_{ics} be the outcome of child i in classroom c and preschool center s. The model to estimate is:

$$Y_{ics} = \beta_0 + \beta_1 T1_{ics} + \beta_2 T2_{ics} + \alpha X_{ics} + \gamma_{cs} + \varepsilon_{ics} \tag{3.1}$$

where $T1_{ics}$ and $T2_{ics}$ indicate assignment to MKit and MKit + present bias groups, respectively; X_{ics} is a vector of baseline characteristics; γ_{cs} is classroom fixed effects, given that the randomization was conducted within classrooms;[15] and ε_{ics} is the error term. Our main identifying assumption is that the treatments are uncorrelated with unobserved determinants of the outcome. The parameters of interest, β_1 and β_2, estimate the intent-to-treat (ITT) effect of being offered the treatments.

The vector of baseline characteristics, X_{ics}, includes child's baseline test scores, which we control for in all specifications due to imbalance in randomization. It also includes sex, age, whether the consent form was completed in Spanish, and indicators for invalid and imputed test scores. These variables are available for all children, even if their parents did not take the baseline survey.

We test for heterogeneous impacts by interacting various baseline characteristics with the treatment indicators. We prespecified the following variables in the analysis plan: (i)

15. We use classroom assignment at the time of randomization. 1 percent of children moved to a different classroom when the baseline assessment was administered, and 3 percent of children were in a different classroom at the time of the first follow-up assessment.

child's baseline test score, (ii) child's sex, (iii) child's race and ethnicity, and (iv) parental present bias. For baseline test scores, we divide the sample into three to investigate whether children with low or high math skills benefit from the intervention. The first group includes children whose scores are below the 25th percentile, the second group—the reference group—includes children between the 25th and 75th percentile in the test score distribution, and the third group is composed of high-achieving children whose test scores are above the 75th percentile.Child's sex indicator divides the sample into males and females. With regard to child's race and ethnicity, our main sample is mainly Hispanic and black. As a consequence, we are not able to look at differences between minority and nonminority students. Nevertheless, we report heterogeneity by race (black and non-black) and language (Spanish consent form and English consent form). For our present bias measure, we divide the sample into two. High present-bias indicator is equal to one if the estimated discount factor for future payoffs is below the median.[16]

The coefficients ϕ_1 and ϕ_2 tell us whether certain types of children benefit more from the MKit and MKit + present bias treatments, respectively.

3.4.2 Treatment effects on the math home environment

While we are primarily interested in estimating impacts on children's math skills, we first estimate the effect of the treatment on potential mechanisms through which the MKit and MKit + present bias interventions may have impacted children's learning. We use the same specification of equation 3.1 to estimate the impact on the math home environment indexes. Data for these indexes come from the follow-up parent survey; thus, the number of observations is 1,040.

Table 3.3 shows treatment impacts on the indexes of barriers, math activities, reading ac-

16. The dummy variables do not necessarily divide the sample into two equal-size groups because the discount factors are only estimated when responses to the Money Game have variation. 50 percent of the responses to the money task have no variation, i.e. respondents chose the same answer throughout the task. Parents who always preferred to take the money at an early date were classified as highly present oriented.

tivities, material investment, and math time. For each index, the first specification includes classroom fixed effects and the second specification in addition controls for child characteristics. The MKit and MKit + present bias interventions did not change parents' reports of the barriers they face to do math activities with their children (columns 1 and 2). Providing the MKit to parents had a positive but not statistically significant effect on parents reported math activities. But the MKit + present bias treatment increased by 0.22σ ($p = 0.05$) the index (columns 3 and 4). The MKit + present bias intervention increased the time that parents spend with their children in reading activities (0.17σ, $p = 0.10$), suggesting the potential for positive spillovers (columns 5 and 6). Both treatments increased by 0.27σ-0.31σ parents' reports of their material investments in their child's math skills (columns 7 and 8). Only the behavioral treatment increased the index of math time, and the effect is statistically significant (columns 9 and 10).

Table 3.3: Treatment Effects on the Math Home Environment

	Ind. neg. barriers		Ind. math act.		Ind. reading act.		Ind. material invest.		Ind. math time	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
MKit	0.06	0.07	0.11	0.12	0.11	0.11	0.30***	0.31***	0.02	0.02
	(0.08)	(0.08)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)
MKit + present bias	-0.02	-0.02	0.22**	0.22**	0.18*	0.17*	0.27***	0.27***	0.24**	0.24**
	(0.08)	(0.08)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)
Control mean	.04	.04	-.15	-.15	-.14	-.14	-.17	-.17	-.15	-.15
Classroom f.e.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Child covariates		Yes		Yes		Yes		Yes		Yes
R2	.23	.24	.24	.25	.23	.24	.2	.2	.33	.33
N	1040	1040	1040	1040	1040	1040	1040	1040	1040	1040

Notes: Robust standard errors reported in parentheses. Estimates are adjusted for child baseline covariates and classroom fixed effects depending on the specification. Dependent variables are indexes constructed via principal component analysis of survey responses, as described in Section 3.2.4. Indexes are standardized to have zero mean and unit standard deviation. Asterisks denote significance level: * $p < .10$, ** $p < .05$, *** $p < .01$.

Taken together, these results suggest that simply providing the MKit to parents had little effect on parents' reported investments in their child's math skills but when the materials are combined with the behavioral tools, parents report more engagement in both math and reading activities and more material investment in these activities. We rely on parents' reports of their activities, which means that measures may not capture actual parental

behavior but may instead reflect a change in parents' perception of their own behavior—that is, that they are doing more because the messaging increases their focus on educational activities. It is also possible that the behavioral tools increase the normative salience of educational activities, resulting in more parents reporting these activities but not necessarily engaging in them. We next examine whether the reported increase in parental engagement translated into higher children's math skills.

3.4.3 Treatment effects on children's early math skills

Table 3.4 shows estimates of equation 3.1 with the WJ as the dependent variable. Each column is a different specification. Columns with odd numbers regress WJ on treatment indicators and classroom fixed effects. Columns with even numbers, in addition, include the set of child covariates as explanatory variables.

Table 3.4: Treatment Effects on WJ Z-Score

	Main sample		No outliers		No imputation		Valid scores		No peer effects	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
MKit	-0.13*	-0.13*	-0.11	-0.11*	-0.16**	-0.16**	-0.11	-0.11	-0.11	-0.12
	(0.07)	(0.07)	(0.07)	(0.07)	(0.08)	(0.08)	(0.08)	(0.08)	(0.09)	(0.09)
MKit + present bias	-0.12*	-0.11	-0.09	-0.08	-0.11	-0.11	-0.10	-0.09	-0.08	-0.08
	(0.07)	(0.07)	(0.07)	(0.07)	(0.08)	(0.08)	(0.08)	(0.08)	(0.09)	(0.09)
Control mean	-.89	-.89	-.88	-.88	-.88	-.88	-.85	-.85	-.9	-.9
Classroom f.e.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes		
Child covariates		Yes		Yes		Yes		Yes		Yes
School f.e.									Yes	Yes
R2	.44	.44	.42	.42	.5	.5	.53	.53	.36	.37
N	1412	1412	1367	1367	1199	1199	1133	1133	1132	1132

Notes: Robust standard errors reported in parentheses. Dependent variable is the Woodcock-Johnson Applied Problems adjusted by age and norm referenced. Estimates are adjusted for child baseline covariates and classroom fixed effects depending on the specification. The sample used to estimate the treatment effects varies across specifications. Columns 1 and 2 are the main sample. Columns 3 and 4 exclude children whose test scores are below the 5th percentile and above the 95th percentile. Columns 5 and 6 exclude test scores that were missing but we imputed. Valid scores sample in columns 7 and 8 excludes children who were identified as having communication difficulties or cognitive disabilities. No peer effects sample in columns 9 and 10 excludes control children in treated classrooms, thus the reference group is control children in untreated classrooms. This specification replaces classroom fixed effects for school fixed effects due to the across-classroom comparison. Asterisks denote significance level: * $p < .10$, ** $p < .05$, *** $p < .01$.

If we focus on columns 1 and 2 of Table 3.4, which show estimates with the main sam-

ple, we see that simply providing the MKit treatment to parents resulted in a marginally significant ($p = 0.10$) negative effect on test scores in both specifications in Table 3.4. The coefficient of the MKit + present bias group is negative, marginally significant in specification without covariates, but is not statistically different from the MKit coefficient. Because we control for baseline math skills, the interpretation of the treatment coefficients is in growth terms. For instance, column 2 indicates that being assigned to the present bias group decreased the growth in math skills by 0.11σ. This does not mean that the math scores for this treatment group were lower at the end of the intervention than at the beginning.

We test the robustness of the results using different ways of handling test score data. First, we exclude very small or very large scores that fall below the 1st percentile or above the 99th percentile of the test score distribution. Second, we drop imputed test scores. Third, we restrict the sample to children with non-missing valid test scores. Imputed and invalid test scores are excluded. These results are in columns 3 through 8 in Table 3.4. For each subsample, we show estimates with and without covariates. Excluding the bottom and top 1 percent reduces the magnitude of the estimates, but the coefficients remain negative and marginally significant ($p = 0.10$) for the MKit only group (see columns 3 and 4). Excluding the imputed scores makes the MKit only coefficient more negative (-0.16) and is statistically significant at 5 percent level. The coefficient on the MKit + present bias intervention is of similar magnitude as the estimate in the whole sample (see columns 5 and 6). Finally, excluding invalid and imputed scores reduces the sizes of the coefficients and we cannot reject the hypothesis that the treatment coefficients are equal to zero (columns 5 and 6).

From these results we cannot conclude that neither the MKit treatment nor the MKit + present bias treatment increased growth in test scores, at least over the short run, despite parents reporting more engagement in math activities with their children from the MKit + present bias treatment. Whether the MKit + present bias treatment degreased growth in scores depends on the model specification.

In summary, the results show that providing materials in the form of a math activity

booklet did not increase children's math skills in the short term. The effect of sending text messages intended to overcome present bias had the opposite effect that we hoped, despite parents reporting higher engagement with their children. Is there any subgroup of children who benefited from the intervention? We respond this question in the next section.

3.4.4 Heterogeneity in short-term treatment effects

In Table 3.5 we examine whether the interventions had differential effects on children's early math skills depending on their family background. We consider differential effects by baseline test scores, gender, race, Spanish language, and present bias. We find a negative differential effect on male children in the MKit group because they scored 0.27σ ($p = 0.05$) lower than female children in this group. We also find a positive and marginally significant effect of the MKit + present bias treatment for children whose parents completed the consent form in Spanish relative to those whose parents did not (0.22σ, $p = 0.10$). However, the total effect for this subgroup (0.04σ = -0.20 - 0.03 + 0.27) is not statistically significant. We find no statistically significant impact by level of parents' impatience.

With our data, we can answer whether parents who opted in to participate were differentially influenced by the treatments. We implemented opt-in and opt-out strategies to enroll parents in MPACT and this difference in recruitment may have drawn different types of parents. Parents in opt-in schools, who make up 58 percent of MPACT participants, actively enrolled in the MPACT program, whereas parents in opt-out schools were automatically enrolled. We test whether treatment parents in opt-in centers were differentially affected by interacting the opt-in indicator with treatment indicators. Because the opt-in indicator is at the preschool level, this variable is absorbed by the classroom fixed effects.

Column 6 of Table 3.5 shows that the impact of the MKit treatment is more negative for children in opt-in centers (-0.34σ, $p = 0.05$) than those in opt-out centers. The effect of the MKit + present bias group is also negative (-0.14σ) but not distinguishable from zero. This means that the math skills of treated children in opt-in centers are growing slower than

Table 3.5: Treatment Effects on WJ Z-Score by Baseline Test Score, Sex, Race, and Impatience

	WJ z-score					
	(1)	(2)	(3)	(4)	(5)	(6)
MKit	-0.09	-0.01	-0.04	-0.13*	-0.15	0.07
	(0.11)	(0.09)	(0.10)	(0.08)	(0.24)	(0.11)
MKit + present bias	-0.11	-0.14	-0.07	-0.20**	0.10	-0.02
	(0.12)	(0.09)	(0.10)	(0.08)	(0.21)	(0.11)
Low WJ z-score	-0.01					
	(0.18)					
High WJ z-score	-0.15					
	(0.11)					
MKit × Low WJ z-score	-0.13					
	(0.18)					
MKit × High WJ z-score	-0.02					
	(0.15)					
MKit + present bias × Low WJ z-score	-0.18					
	(0.20)					
MKit + present bias × High WJ z-score	0.12					
	(0.15)					
Male		0.00				
		(0.08)				
MKit × Male		-0.27**				
		(0.14)				
MKit + present bias × Male		0.06				
		(0.14)				
Black			-0.19			
			(0.13)			
MKit × Black			-0.12			
			(0.16)			
MKit + present bias × Black			0.08			
			(0.17)			
Spanish consent				-0.03		
				(0.10)		
MKit × Spanish consent				0.00		
				(0.15)		
MKit + present bias × Spanish consent				0.27*		
				(0.14)		
Impatient					0.09	
					(0.16)	
MKit × Impatient					0.10	
					(0.26)	
MKit + present bias × Impatient					-0.25	
					(0.26)	
Opt-in						0.00
						(.)
MKit × Opt-in						-0.34**
						(0.14)
MKit + present bias × Opt-in						-0.14
						(0.14)
Classroom f.e.	Yes	Yes	Yes	Yes	Yes	Yes
Child covariates	Yes	Yes	Yes	Yes	Yes	Yes
R2	.45	.45	.47	.45	.53	.45
N	1412	1412	1098	1412	606	1412

Notes: Robust standard errors reported in parentheses. Estimates are adjusted for WJ at baseline, child baseline covariates and classroom fixed effects. Dependent variable is the Woodcock Johnson IV Applied Problems adjusted by age and norm referenced. Low and high WJ z-score indicates whether baseline WJ z-score in the bottom or top quartiles, respectively. Male is child's sex. Black indicates whether the child's parent identifies as non-Hispanic African-American. Spanish consent is equal to one if the consent form was filed in Spanish. The impatience indicator is calculated with the time-preference task for money, as described in Section 3.2.4. Opt-in indicator takes value of one if preschool center implemented an opt-in strategy to enroll in MPACT. Asterisks denote significance level: * $p < .10$, ** $p < .05$, *** $p < .01$.

those of control children. There are differences between families in opt-in and opt-out centers (e.g., opt-out centers have more Spanish-speaking families); as a result, we cannot conclude that the different responses to the treatment are due to differences in parents' motivation to participate in MPACT or the contexts where children learn.

In sum, our heterogeneity analysis provides no strong evidence of positive effects by the prespecified child and parental characteristics. Spanish-speaking families seem to benefit from the MKit + present bias group, but the estimated effects have large standard errors. We find differential negative effects of the MKit treatment on male children and children in opt-in centers, but here again standard errors are about 40 percent of point estimates. We discuss potential possible explanations for these findings in the next section.

3.5 Why Are There No Positive Treatment Impacts in the Short Term?

In this section we consider possible explanations for the negative and sometimes null effects of the MKit and MKit + present bias groups.

3.5.1 Did parents receive the treatment they were supposed to receive?

If parents did not receive the materials and text messages that they were supposed to receive, then they were not treated. We distributed the physical materials (reading books for the control group and MKits for the treatment groups) in person to make sure the guardians of the participant children received the correct material. If a parent was not able to pick up her materials, we left them at their children's cubicle and verified with preschool staff that the materials had been picked up. For the text messages, we used a system that reports back invalid phone numbers, which in our sample was only 1 percent. Our material distribution method suggests fairly high fidelity to the treatment protocol. In addition, it is not clear that not receiving the treatment materials ought to result in slower growth in test scores

compared to the control group.

3.5.2 Did control students benefit from having treated peers?

One possible explanation for the lack of benefit from these treatments is the possibility that children in the control group benefited from having treated peers. This would dilute the difference between treated and control children. Students within classrooms interact with each other, and this may cause spillovers between control and treated participants. Spillover effects may also arise due to teacher responses to having treated children in the classroom. Teachers in classrooms with treated children may change their teaching materials, or children in these classrooms may be less disruptive children. We exploit our two-stage randomization to test for peer effects, as children in untreated classrooms do not have treated peers.

Columns 9 and 10 of Table 3.4 show our test for peer effects. We drop children in the control group in treated classrooms; hence, the comparison is between treated children in treated classrooms and control children in untreated classrooms. The latter group serves as the control purged of peer effects. Because these group of children are in different classrooms, we replace classroom fixed effects with site fixed effects. The treatment remain negative around -0.12 for the MKit group and -0.08 for the MKit + present bias group, which indicates there is no evidence of spillover effects.

3.5.3 Did the treatment increase math anxiety?

If the treatment made parents more anxious about math, it is possible that interactions with their children on math topics might have become more stressful and confusing, resulting in less math growth for the child (even if the parent spent more time in math activities). The correlation between follow-up WJ and math anxiety score is indistinguishable from zero ($\rho = -0.02$), which indicates that math anxiety is not a likely reason for the negative treatment effects. More evidence of this is the no significant treatment effects on math anxiety (see columns 1 and 2 of Table 3.6). The MKit + present bias group reported higher

math anxiety than the control group at the end of the intervention, although the effects are not statistically significant at 5 percent level.

Table 3.6: Treatment Effects on Parental Math Anxiety and Beliefs

	Math anxiety		Growth mindset		Marginal effect	
	(1)	(2)	(3)	(4)	(5)	(6)
MKit	-0.11	-0.15	-0.07	-0.07	0.04	0.05
	(0.27)	(0.27)	(0.10)	(0.10)	(0.11)	(0.11)
MKit + present bias	0.14	0.12	-0.06	-0.07	-0.03	-0.03
	(0.27)	(0.27)	(0.10)	(0.10)	(0.10)	(0.10)
Control mean	2.5	2.5	4.4	4.4	-.03	-.03
Classroom f.e.	Yes	Yes	Yes	Yes	Yes	Yes
Child covariates		Yes		Yes		Yes
R2	.19	.2	.25	.27	.16	.17
N	1040	1040	1040	1040	1040	1040

Notes: Robust standard errors reported in parentheses. Estimates are adjusted for child baseline covariates and classroom fixed effects depending on the specification. Math anxiety score comes from the question "How nervous does math make you?" Growth mindset is calculated using the Dweck scale. Marginal effect is computed as the increase in a child's ranking of her math skills if her parent were to spend 15 more minutes in education activities. The impatience indicator is calculated with the time-preference task for money, as described in Section 3.2.4. Asterisks denote significance level: * $p < .10$, ** $p < .05$, *** $p < .01$.

3.5.4 Did treatment change parental beliefs from a growth to a fixed mindset?

It is possible that changes in growth mindset result in changes in test scores. But the correlation between test scores and parents' growth mindset is 0.08 ($p < 0.01$), so this is unlikely to explain the negative effect of MKit + present bias. Regression analysis supports this. We observe in columns 3 and 4 of Table 3.6 that treatment did not change parents' growth mindset scores.

3.5.5 Did parents in treatment groups lower their expected returns on investment?

Similar to growth mindset, one could imagine that parents realized their investments were less effective if their children did not appear to be learning from the math activities. They may have tried different ways to teach but were not successful. We find no support for this. According to columns 5 and 6 of Table 3.6, there was hardly any change in parental beliefs about the returns on their investments to improve the child's math skills.

3.5.6 Did students hit a ceiling in the WJ test?

If students reached a ceiling in the WJ test, we would find null treatment impacts, even if the treatment improved math skills. To examine this hypothesis, we look at the distribution of the WJ raw score, specifically at the last correct question in the test. The test items are arranged in order of difficulty and the WJ Applied Problems subtest, in particular, has a stopping rule after five items answered incorrectly.[17]

Appendix Figure C.2.1 plots the distribution of the last correct question by treatment group. We see a spike at question number 12. More than 20 percent of children stopped at this question, which is more than double than any other question. Looking at the questionnaire, the items following question 12 were about adding and subtracting. Children in our sample seem to not excel in this skill, and this might be one reason their WJ z-score, which is norm referenced, is below zero. We asked in the parent surveys how frequently they did specific math activities (see *index of math activities* in Section 3.2.4). Helping the child add and subtract is the least frequent activity, with 27 percent of parents declaring never doing it in the past week. The next activity with low frequency is reading books that have

17. The WJ Applied Problems continues page by complete page until the examinee responds incorrectly to five consecutive items and has completed the page. If the examinee answers four items incorrectly and answers the fifth item correctly, the stop rule resets and the test continues. We define the ceiling as the last correct question. The raw score is the sum of all correct questions, and it may not coincide with the last correct question.

number, with 7 percent of parents reporting never doing it. Although we cannot make causal statements, the low level of parental engagement in adding and subtracting activities may contribute to the low skill-level of the child in those areas.

3.5.7 Other hypotheses

Other hypotheses may explain the null and often negative short-term treatment impacts. For example, disadvantaged parents may not have the necessary knowledge or skills to teach math to their children, even with the MKit. The MKit may have been confusing for the parents to understand, it could have been not culturally appropriate, or the treatments may not have produced changes in behavior strong enough to matter for children's learning. It could also be the case that it takes longer time for the effects to be realized. Although we cannot test all imaginable hypotheses with the data we have now, we can answer some of them with the data that will be collected at six and 12 months postintervention. There is a longer period for the effects to be realized. In addition, parents will receive booster text messages to encourage them to continue interacting with their children. We will have a more definitive answer as to whether the behavioral intervention worked in the future.

3.6 Conclusions

Short-term treatment impacts show an increase in the math environment at home but no significant treatment impacts on test score outcomes. The opposite effects are observed. The MKit and MKit + present bias interventions seem to have slowed the acquisition of math skills on average. However, depending on how we handle test score data, the effects are not distinguishable from zero. Subgroup analyses show that children with Spanish-speaking families benefit the most from the intervention, but these estimates have high standard errors.

Our study makes three major contributions to the literature. First, to the best of our

knowledge, this is one of the few (if not the first) randomized evaluation that uses behavioral tools to increase early math skills. Second, we implemented a two-stage randomization in a preschool setting to investigate peer effects. Third, we provide evidence that present bias for money and effort are not positively correlated.

The results of the short-term impacts are not as encouraging as we hoped, but we will have to wait until the long-term outcomes are collected to have a definitive result. The negative—and sometimes null—short-term treatment effects may indicate that parents are incorporating new behaviors and habits that are not effective (at first) to increase children's math skills, but with time they learn to be more effective. It could be possible that the benefits of incorporating new habits manifest in the long run.

REFERENCES

Aaronson, D., Barrow, L., and Sander, W. (2007). Teachers and student achievement in the chicago public high schools. *Journal of labor Economics*, 25(1):95–135.

Aikens, N., Klein, A. K., Knas, E., Reid, M., Esposito, A. M., Manley, M., Malone, L., Tarullo, L., Lukashanets, S., and West, J. (2017). Descriptive data on head start children and families from faces 2014: Fall 2014 data tables and study design. Technical Report OPRE Report 2017-97, Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services., Washington, DC.

Andreoni, J., Kuhn, M. A., and Sprenger, C. (2015). Measuring time preferences: A comparison of experimental methods. *Journal of Economic Behavior & Organization*, 116:451–464.

Andreoni, J. and Sprenger, C. (2012). Estimating time preferences from convex budgets. *American Economic Review*, 102(7):3333–56.

Aucejo, E. M. (2011). Assessing the role of teacher-student interactions. *Manuscript, Duke University*.

Bacher-Hicks, A., Kane, T. J., and Staiger, D. O. (2014). Validating teacher effect estimates using changes in teacher assignments in los angeles. Technical report, National Bureau of Economic Research.

Baroody, A. J., Eiland, M. D., Purpura, D. J., and Reid, E. E. (2012). Fostering at-risk kindergarten children's number sense. *Cognition and Instruction*, 30(4):435–470.

Becker, G. S. and Tomes, N. (1979). An equilibrium theory of the distribution of income and intergenerational mobility. *Journal of political Economy*, 87(6):1153–1189.

Bell, A., Chetty, R., Jaravel, X., Petkova, N., and Van Reenen, J. (2018). Who Becomes an Inventor in America? The Importance of Exposure to Innovation*. *The Quarterly Journal of Economics*, 134(2):647–713.

Berkowitz, T., Schaeffer, M. W., Maloney, E. A., Peterson, L., Gregor, C., Levine, S. C., and Beilock, S. L. (2015). Math at home adds up to achievement in school. *Science*, 350(6257):196–198.

Blackwell, L. S., Trzesniewski, K. H., and Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child development*, 78(1):246–263.

Blazar, D. and Kraft, M. A. (2017). Teacher and teaching effects on students' attitudes and behaviors. *Educational evaluation and policy analysis*, 39(1):146–170.

Borman, G. D. and Kimball, S. M. (2005). Teacher quality and educational equality: Do teachers with higher standards-based evaluation ratings close student achievement gaps? *The elementary school journal*, 106(1):3–20.

Boyd, D., Grossman, P., Ing, M., Lankford, H., Loeb, S., and Wyckoff, J. (2011). The influence of school administrators on teacher retention decisions. *American Educational Research Journal*, 48(2):303–333.

Bryan, G., Karlan, D., and Nelson, S. (2010). Commitment devices. *Annual Review of Economics*, 2(1):671–698.

Burks, S., Carpenter, J., Götte, L., and Rustichini, A. (2012). Which measures of time preference best predict outcomes: Evidence from a large-scale field experiment. *Journal of Economic Behavior and Organization*, 84(1):308–320.

Campbell, S. L. and Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal*, 55(6):1233–1267.

Carvalho, L. S., Meier, S., and Wang, S. W. (2016). Poverty and economic decision-making: Evidence from changes in financial resources at payday. *The American Eonomic Review*, 106(2):260–284.

Chaplin, D., Gill, B., Thompkins, A., and Miller, H. (2014). Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in the pittsburgh public schools. rel 2014-024. *Regional Educational Laboratory Mid-Atlantic*.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014a). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9):2593–2632.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014b). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review*, 104(9):2633–79.

Claessens, A. and Engel, M. (2013). How important is where you start? early mathematics knowledge and later school success. *Teachers College Record*, 115(6).

Claro, S., Paunesku, D., and Dweck, C. S. (2016). Growth mindset tempers the effects of poverty on academic achievement. *Proceedings of the National Academy of Sciences*, 113(31):8664–8668.

Clotfelter, C. T., Ladd, H. F., and Vigdor, J. L. (2006a). Teacher-student matching and the assessment of teacher effectiveness. *Journal of human Resources*, 41(4):778–820.

Clotfelter, C. T., Vigdor, J. L., and Ladd, H. F. (2006b). Federal oversight, local control, and the specter of "resegregation" in southern schools. *American Law and Economics Review*, 8(2):347–389.

Dee, T. S. (2004). Teachers, race, and student achievement in a randomized experiment. *The Review of Economics and Statistics*, 86(1):195–210.

Del Boca, D., Flinn, C., and Wiswall, M. (2013). Household Choices and Child Development. *The Review of Economic Studies*, 81(1):137–185.

DellaVigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic Literature*, 47(2):315–72.

DellaVigna, S. and Malmendier, U. (2006). Paying not to go to the gym. *American Economic Review*, 96(3):694–719.

DellaVigna, S. and Paserman, M. D. (2005). Job search and impatience. *Journal of Labor Economics*, 23(3):527–588.

Doherty, K. and Jacobs, S. (2013). Connect the dots: Using evaluations of teacher effectiveness to inform policy and practice. national council on teacher quality.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., et al. (2007). School readiness and later achievement. *Developmental psychology*, 43(6):1428.

Duncan, G. J. and Magnuson, K. (2013). Investing in preschool programs. *Journal of Economic Perspectives*, 27(2):109–32.

Egalite, A. J. and Kisida, B. (2018). The effects of teacher match on students' academic perceptions and attitudes. *Educational Evaluation and Policy Analysis*, 40(1):59–81.

Egalite, A. J., Kisida, B., and Winters, M. A. (2015). Representation in the classroom: The effect of own-race teachers on student achievement. *Economics of Education Review*, 45:44 – 52.

Fiorini, M. and Keane, M. P. (2014). How the allocation of children's time affects cognitive and noncognitive development. *Journal of Labor Economics*, 32(4):787–836.

Fryer, R. G. and Levitt, S. D. (2006). The black-white test score gap through third grade. *American Law and Economics Review*, 8(2):249–281.

Fryer, R. G. and Levitt, S. D. (2013). Testing for racial differences in the mental ability of young children. *American Economic Review*, 103(2):981–1005.

Gershenson, S. (2016). Linking teacher quality, student attendance, and student achievement. *Education Finance and Policy*, 11(2):125–149.

Goldhaber, D., Lavery, L., and Theobald, R. (2015). Uneven playing field? assessing the teacher quality gap between advantaged and disadvantaged students. *Educational researcher*, 44(5):293–307.

Guryan, J., Hurst, E., and Kearney, M. (2008). Parental education and parental time with children. *Journal of Economic perspectives*, 22(3):23–46.

Harackiewicz, J. M., Rozek, C. S., Hulleman, C. S., and Hyde, J. S. (2012). Helping parents to motivate adolescents in mathematics and science: An experimental test of a utility-value intervention. *Psychological Science*, 23(8):899–906. PMID: 22760887.

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, 312(5782):1900–1902.

Heckman, J. J. (2008). Schools, skills, and synapses. *Economic inquiry*, 46(3):289–324.

Hsin, A. and Felfe, C. (2014). When does time matter? maternal employment, children's time with parents, and child development. *Demography*, 51(5):1867–1894.

Jackson, C. K. (2018). What do test scores miss? the importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, 126(5):2072–2107.

Jacob, B. A. and Lefgren, L. (2008). Can principals identify effective teachers? evidence on subjective performance evaluation in education. *Journal of labor Economics*, 26(1):101–136.

Jones, D. and Mahajan, A. (2015). Time-inconsistency and saving: Experimental evidence from low-income tax filers. *National Bureau of Economic Research*, (21272).

Kalil, A., Ryan, R., and Corey, M. (2012). Diverging destinies: Maternal education and the developmental gradient in time with children. *Demography*, 49(4):1361–1383.

Kalogrides, D. and Loeb, S. (2013). Different teachers, different peers: The magnitude of student sorting within schools. *Educational Researcher*, 42(6):304–316.

Kane, T. J., McCaffrey, D. F., Miller, T., and Staiger, D. O. (2013). Have we identified effective teachers? validating measures of effective teaching using random assignment. In *Research Paper. MET Project. Bill and Melinda Gates Foundation.*

Kane, T. J. and Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Technical report, National Bureau of Economic Research.

Kane, T. J. and Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. In *Research Paper. MET Project. Bill and Melinda Gates Foundation.*

Koedel, C. (2008). Teacher quality and dropout outcomes in a large, urban school district. *Journal of Urban Economics*, 64(3):560 – 572.

Konstantopoulos, S. (2009). Effects of teachers on minority and disadvantaged students' achievement in the early grades. *The Elementary School Journal*, 110(1):92–113.

Kraft, M. A. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources*, 54(1):1–36.

Kraft, M. A. and Papay, J. P. (2014). Can professional environments in schools promote teacher development? explaining heterogeneity in returns to teaching experience. *Educational evaluation and policy analysis*, 36(4):476–500.

Ladd, H. F. and Sorensen, L. C. (2017). Returns to teacher experience: Student achievement and motivation in middle school. *Education Finance and Policy*, 12(2):241–279.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2):443–478.

Lankford, H., Loeb, S., and Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational evaluation and policy analysis*, 24(1):37–62.

Lefgren, L. and Sims, D. (2012). Using subject test scores efficiently to predict teacher value-added. *Educational Evaluation and Policy Analysis*, 34(1):109–121.

Lockwood, J. and McCaffrey, D. F. (2009). Exploring student-teacher interactions in longitudinal achievement data. *Education finance and policy*, 4(4):439–467.

Loeb, S., Soland, J., and Fox, L. (2014). Is a good teacher a good teacher for all? comparing value-added of teachers with their english learners and non-english learners. *Educational Evaluation and Policy Analysis*, 36(4):457–475.

Lutz, B. (2011). The end of court-ordered desegregation. *American Economic Journal: Economic Policy*, 3(2):130–68.

Magnuson, K. and Duncan, G. J. (2016). Can early childhood interventions decrease inequality of economic opportunity? *RSF: The Russell Sage Foundation Journal of the Social Sciences*, 2(2):123–141.

Maloney, E. A., Ramirez, G., Gunderson, E. A., Levine, S. C., and Beilock, S. L. (2015). Intergenerational effects of parents' math anxiety on children's math achievement and anxiety. *Psychological Science*, 26(9):1480–1488.

Mayer, S. E., Kalil, A., Oreopoulos, P., and Gallegos, S. (2018). Using behavioral insights to increase parental engagement: The parents and children together intervention. *Journal of Human Resources*, pages 0617–8835R.

Meier, S. and Sprenger, C. (2010). Present-biased preferences and credit card borrowing. *American Economic Journal: Applied Economics*, 2(1):193–210.

Meier, S. and Sprenger, C. D. (2015). Temporal stability of time preferences. *Review of Economics and Statistics*, 97(2):273–286.

Milner IV, H. R. (2006). The promise of black teachers' success with black students. *Educational Foundations*, 20:89–104.

Moiduddin, E., Bush, C., Manley, M., Aikens, N., Tarullo, L., Malone, L., and Lukashanets, S. (2017). A portrait of head start classrooms and programs in spring 2015: Faces 2014-2015 data tables and study design. Technical Report OPRE Report 2017-101, Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services., Washington, DC.

Nguyen, T., Watts, T. W., Duncan, G. J., Clements, D. H., Sarama, J. S., Wolfe, C., and Spitler, M. E. (2016). Which preschool mathematics competencies are most predictive of fifth grade achievement? *Early childhood research quarterly*, 36:550–560.

O'Donoghue, T. and Rabin, M. (1999). Doing it now or later. *American Economic Review*, 89(1):103–124.

Office of Head Start (2017). PIR Reports. Technical report.

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*, 48(1):163–193.

Papay, J. P. and Kraft, M. A. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics*, 130:105–119.

Paufler, N. A. and Amrein-Beardsley, A. (2014). The random assignment of students into elementary classrooms: Implications for value-added analyses and interpretations. *American Educational Research Journal*, 51(2):328–362.

Petek, N. and Pope, N. (2016). The multidimensional impact of teachers on students. Technical report, University of Chicago Working Paper.

Purpura, D. J. and Lonigan, C. J. (2015). Early numeracy assessment: The development of the preschool early numeracy scales. *Early education and development*, 26(2):286–313.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*, volume 1. Sage.

Reardon, S. F. and Portilla, X. A. (2016). Recent trends in income, racial, and ethnic school readiness gaps at kindergarten entry. *AERA Open*, 2(3):2332858416657343.

Redding, C. (2019). A teacher like me: A review of the effect of student-teacher racial/ethnic matching on teacher perceptions of students and student academic and behavioral outcomes. *Review of Educational Research*, 0(0).

Rivkin, S. G., Hanushek, E. A., and Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2):417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American economic review*, 94(2):247–252.

Rockoff, J. E. and Speroni, C. (2011). Subjective and objective evaluations of teacher effectiveness: Evidence from new york city. *Labour Economics*, 18(5):687–696.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education finance and policy*, 4(4):537–571.

Rothstein, J. (2017). Measuring the impacts of teachers: comment. *American Economic Review*, 107(6):1656–1684.

Sanders, W. L. and Horn, S. P. (1998). Research findings from the tennessee value-added assessment system (tvaas) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, 12(3):247–256.

Sartain, L., Stoelinga, S. R., and Brown, E. R. (2011). *Rethinking Teacher Evaluation in Chicago: Lessons Learned from Classroom Observations, Principal-Teacher Conferences, and District Implementation. Research Report.* ERIC.

Sass, T. R., Hannaway, J., Xu, Z., Figlio, D. N., and Feng, L. (2012). Value added of teachers in high-poverty schools and lower poverty schools. *Journal of urban Economics*, 72(2-3):104–122.

School, C. P. (2019). Reach students: Teacher performance evaluation.

Sporte, S. E. and Jiang, J. Y. (2016). *Teacher Evaluation in Practice: Year 3 Teacher and Administrator Perceptions of REACH. Research Brief.* ERIC.

Steinberg, M. P. and Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-nclb era. *Education Finance and Policy*, 11(3):340–359.

Steinberg, M. P. and Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis*, 38(2):293–317.

Watts, T. W., Duncan, G. J., Clements, D. H., and Sarama, J. (2018). What is the long-run impact of learning mathematics during preschool? *Child development*, 89(2):539–555.

Whitehurst, G., Chingos, M. M., and Lindquist, K. M. (2014). Evaluating teachers with classroom observations. *Brown Center on Education Policy: Brookings Institute.*

York, B. N., Loeb, S., and Doss, C. (2018). One step at a time: The effects of an early literacy text messaging program for parents of preschoolers. *Journal of Human Resources*, pages 0517–8756R.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association*, 57(298):348–368.

APPENDIX A

APPENDIX TO CHAPTER 1 (TEACHERS' COMPARATIVE ADVANTAGE)

A.1 Reliability Weights

Starting with Eq. 1.17, the elements of the matrixes $\mathbf{\Gamma}$ and $\boldsymbol{\gamma}_k$ are function of second moment conditions (variances and covariances of parameters). Under the stationarity assumptions, these elements are:

$$
\begin{aligned}
\left[\boldsymbol{\phi}_{k,k}\right]_s &= Cov(\bar{A}_{jkt}, \bar{A}_{jks}) &&= \sigma_{\mu_k|t-s|} \\
\left[\boldsymbol{\phi}_{k,m}\right]_s &= Cov(\bar{A}_{jkt}, \bar{A}_{jms}) &&= \sigma_{\mu_k\mu_m|t-s|} \\
\left[\mathbf{\Gamma}_{mm}\right]_{ss} &= Var(\bar{A}_{jms}) &&= \sigma^2_{\mu_m} + \sigma^2_{\theta_m} + \frac{\sigma^2_\epsilon}{n_{jms}} \\
\left[\mathbf{\Gamma}_{mm}\right]_{ss'} &= Cov(\bar{A}_{jms}, \bar{A}_{jms'}) &&= \sigma_{\mu_m|s-s'|} \\
\left[\mathbf{\Gamma}_{mn}\right]_{ss} &= Cov(\bar{A}_{jms}, \bar{A}_{jns}) &&= \sigma_{\mu_m\mu_n} + \sigma_{\theta_m\theta_n} \\
\left[\mathbf{\Gamma}_{mn}\right]_{ss'} &= Cov(\bar{A}_{jms}, \bar{A}_{jns'}) &&= \sigma_{\mu_m\mu_n|s-s'|} \\
\left[\mathbf{\Gamma}_{mn}\right]_{s's} &= Cov(\bar{A}_{jms'}, \bar{A}_{jns}) &&= \sigma_{\mu_m|s-s'|\mu_n}
\end{aligned}
$$

for all k, $n \neq m$, and $s' < s$.

A.2 Student-Teacher Racial Match Effects in CPS

In this section we reproduce results of studies on student-racial match, specifically Egalite et al. (2015). With data from the universe of students in Florida public schools, they use a student fixed-effect model to estimate the relationship between student-teacher racial match and student achievement. We follow their procedure and estimate:

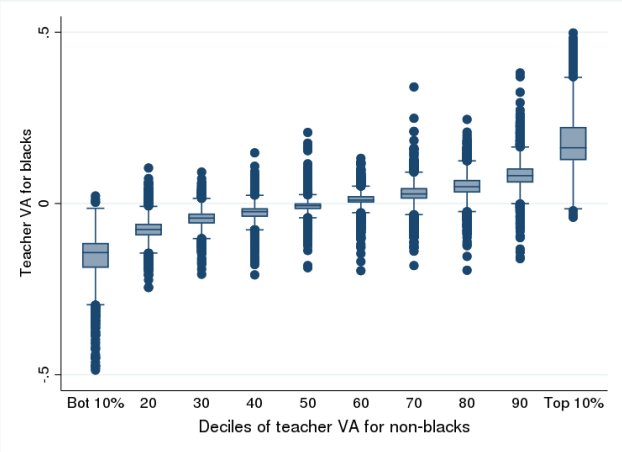
$$A^*_{it} = \beta_0 + X'_{ikt}\beta_1 + X'_{jt}\beta_2 + \gamma RaceMatch_{ijt} + \alpha_i + \varepsilon_{it} \tag{A.1}$$

where A^*_{it} is the test score of student i in time t, who was assigned to teacher j; X'_{it} is a vector of student, classroom and school characteristics (same covariates that we used in Eq. 1.1); $X'_{j(i)t}$ denotes the characteristics of teacher j, including years of experience, experience squared, race, and gender; $RaceMatch_{ij}$ is an indicator variable that takes value one if the student's race matches with that of her teacher; τ_t is year fixed effect; α_i is student fixed effect; and ε_{it} is the error term clustered at the cohort level. To be consistent with our findings, the race variable is either black or non-black. Therefore, $RaceMatch_{ij}$ takes value of one if the student and teacher are both black or non-black. The parameter γ is the parameter of interest, and it estimates the change in test scores when a student is assigned to a teacher who shares same racial characteristics.

Table A.3.2 shows our estimates in Panel A and reports Egalite et al. (2015)'s main results in Panel B. The model is estimated separately for math and reading and, for each subject, elementary and middle schools. Our estimates are remarkably similar to those of Egalite et al. (2015). Being assigned to a same-race teacher increases test scores. The effect is larger for math than reading and, within subject, the effects are larger in elementary schools than middle schools.

A.3 Appendix Figures and Tables

Figure A.3.1: Distribution of Teacher VA on Black Students by Deciles of Teacher VA on Non-Black Students



Notes: The sample is divided into four equally sized groups based on the distribution of non-black VA. Then we construct box plots of black VA within quartiles. The lines inside the box represent the median value and the upper and lower sides of the box are the 25th and 75th percentile. The vertical lines (whiskers) extend two-thirds of the quartile values, and the points are value outside of the whiskers. This figure provides evidence of multivariate teacher effects because the means of each decile do not fall on a straight line and many VA on blacks in lower deciles of VA of non-blacks are higher than some values on higher deciles.

Table A.3.1: Number of Observations Used to Estimate Variance-Covariance Matrix of Teacher VA on Blacks and Non-Blacks

	Math				Reading			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A. Elementary schools								
Var 1:	VA black		VA non-black		VA black		VA non-black	
Var 2:	VA black	VA non-black	VA black	VA non-black	VA black	VA non-black	VA black	VA non-black
Lag of var 2								
1	3129	837	843	6607	3150	821	843	6961
2	1969	534	571	4445	1979	518	586	4720
3	1222	347	389	2938	1229	326	404	3129
4	717	214	252	1806	699	191	251	1935
5	400	113	143	990	376	99	150	1070
6	212	61	89	545	208	57	96	601
7	60	32	36	232	64	26	39	254
Panel B. Middle schools								
Var 1:	VA black		VA non-black		VA black		VA non-black	
Var 2:	VA black	VA non-black	VA black	VA non-black	VA black	VA non-black	VA black	VA non-black
Lag of var 2								
1	1990	910	958	2713	2107	886	887	3428
2	1375	614	726	1924	1321	568	608	2314
3	970	456	575	1420	924	402	453	1625
4	647	291	409	984	556	255	303	1077
5	396	181	282	642	343	164	205	678
6	229	113	181	386	184	104	129	418
7	48	29	50	130	45	39	39	164

Notes: Panel A and B of this table show the number of observations used to estimate the variance-covariance matrix of teacher VA presented in Table 1.2.

Table A.3.2: Estimates of Student-Teacher Racial Match Using Within-Student Variation

	Reading			Math			Both
	All grades	Elementary	Middle	All grades	Elementary	Middle	All grades
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: CPS							
Race match	0.005	0.001	0.013**	0.011**	0.008	0.008	0.009***
	(0.004)	(0.006)	(0.006)	(0.005)	(0.008)	(0.009)	(0.003)
R-squared	0.88	0.93	0.91	0.89	0.92	0.93	0.82
Sample	873,480	432,999	440,481	881,332	441,231	440,101	1,754,812
Panel B: Egalite et al (2015)							
Race match	0.002**	0.004***	-0.001	0.008***	0.014***	0.002**	
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	
R-squared	0.85	0.89	0.88	0.87	0.9	.90	
N	8,598,592	3,142,898	5,455,694	8,729,364	3,090,124	5,639,240	

Notes: Panel A reports estimates of student-teacher racial match in CPS by subject and grade level. These estimates come from student-fixed effect regressions of student test scores (not residualized) and an indicator of racial match. This indicator takes value of 1 if the student and her assigned teacher share the same race (black or non-black). Columns 1-3 uses reading test scores as dependent variable, columns 4-6 uses math test scores, and column 7 pools both test scores. Explanatory variables include those used in estimating teacher VA, as described in Section , and teacher's sex, race/ethnicity (white, black, or Hispanic), experience in CPS, experience squared, and own-type teacher VA. Panel B is taken from Table 3 of Egalite et al. (2015).

APPENDIX B

APPENDIX TO CHAPTER 2 (EVALUATOR BIAS)

Table B.0.1: Description of Student Survey Questions Used to Construct Survey Indexes

Management Index Items
How much do you disagree or agree with the following statement about your [TARGET] class? (Strongly disagree, disagree, agree, or strongly agree) 1. I get distracted from my work by other students acting out in this class. 2. This class is out of control. 3. My classmates do not behave the way my teacher wants them to.
Instruction Index Items
How much do you disagree or agree with the following statements about your teacher in your [TARGET] class? My teacher... (Strongly disagree, Disagree, Agree, or Strongly agree) 1. Often connects what I am learning to life outside of the classroom. 2. Encourages students to share their ideas about things we are studying in class. 3. Often requires me to explain my answers. 4. Encourages us to consider different solutions or points of view. 5. Doesn't let students give up when the work gets hard. *How often does the following occur?* (Very little, Some, Quite a bit, A great deal) 1. In my [TARGET] class, we talk about different solutions or points of view. *How much do you disagree or agree with the following statements about your [TARGET] class?* (Strongly disagree, Disagree, Agree, or Strongly agree) 1. This class really makes me think. 2. I'm really learning a lot in this class. *To what extent do you disagree or agree with the following statements? In my [TARGET] class, my teacher...* (Strongly disagree, Disagree, Agree, or Strongly agree) 1. Expects everyone to work hard. 2. Expects me to do my best all the time. 3. Wants us to become better thinkers, not just memorize things. *In your [TARGET] class, how often...* (Never, Once in a while, Most of the time, All the time) 1. Are you challenged? 2. Do you have to work hard to do well? 3. Does the teacher ask difficult questions on tests? 4. Does the teacher ask difficult questions in class? *How much do you disagree or agree with the following statements about your [TARGET] class?* (Strongly disagree, Disagree, Agree, or Strongly agree) 1. I learn a lot from feedback on my work. 2. It's clear to me what I need to do to get a good grade. 3. The work we do in class is good preparation for the test. 4. The homework assignments help me to learn the course material. 5. I know what my teacher wants me to learn in this class. *How much do you disagree or agree with the following statements about your [TARGET] class?* (Strongly disagree, Disagree, Agree, or Strongly agree) 1. I usually look forward to this class. 2. I work hard to do my best in this class. 3. Sometimes I get so interested in my work I don't want to stop 4. The topics we are studying are interesting and challenging. *How much do you disagree or agree with the following statements about your [TARGET] class? The teacher for this class ...* (Strongly disagree, Disagree, Agree, or Strongly agree) 1. Helps me catch up if I am behind. 2. Is willing to give extra help on schoolwork if I need it. 3. Notices if I have trouble learning something. 4. Gives me specific suggestions about how I can improve my work in this class. 5. Explains things in a different way if I don't understand something in class.

Notes: This table shows the questions from student surveys that we use to contruct the Managemenet and Instruction Indexes.

APPENDIX C

APPENDIX TO CHAPTER 3 (THE MPACT INITIATIVE)

C.1 Estimation of Time-Preference Parameters

In the Money Game, parents were asked to choose between an amount of money that they could receive immediately or an amount that they could receive later. In total, each parent answered 15 of these questions. In an effort to simplify the choice situation, each of the questions offered four choices only (two inner solutions and two corner solutions), as in Andreoni et al. (2015). The first five questions offered payments *now* and *3 weeks* later, the next five questions offered payments *now* and *6 weeks* later, and the last five questions offered payments in *3* and *6 weeks*.

The experimental budget is fixed at $m = \$24$ and five prices are implemented in each choice set, summarized by $P \in \{1, 1.14, 1.33, 1.60, 2\}$. These values, and the 3- and 6-week time horizons, were chosen for consistency with our prior work. Denoting by c_t and c_{t+k} consumptions at a soon and later date, respectively, the budget constraint is

$$Pc_t + c_{t+k} = m \tag{C.1}$$

We use the CTB method (first introduced by Andreoni and Sprenger, 2012) to estimate both a discount rate for parents and their present bias, which have been shown to predict outcomes, such as smoking, leaving a job, and credit score (Burks et al., 2012). We assume that parents have preferences for the experimental payments c in period t and $t + k$ and their utility function can be represented by $U(c_t, c_{t+k})$. U is a time-separable, constant relative risk averse function with a quasi-hyperbolic structure for discounting (Laibson, 1997; O'Donoghue and Rabin, 1999) and is given by

$$U(c_t, c_{t+k}) = c_t^{\alpha} + \beta^{\mathbf{1}[t=0]}\delta^k c_{t+k}^{\alpha} \tag{C.2}$$

where $\mathbf{1}[t = 0]$ is an indicator function that takes value 1 if time t is in the present, β captures the degree of present bias, δ denotes the individual's discount rate, and α denotes the curvature of the utility function.

Parents maximize utility subject to the budget constraint in equation C.1. The first order conditions yield the standard condition of marginal rate of substitution equals the price ration. After substituting in the budget constraint, we get

$$c_t = \frac{m\left(P\beta^{\mathbf{1}[t=0]}\delta^k\right)^{\frac{1}{\alpha-1}}}{1 + P\left(P\beta^{\mathbf{1}[t=0]}\delta^k\right)^{\frac{1}{\alpha-1}}} \tag{C.3}$$

We estimate the demand function with a non-linear regression equation to obtain values for β, δ and α.

Some individuals did not respond to all questions or had choices that violate the law of demand. Our analysis excluded individuals with any incomplete answers. To increase sample size, however, we manually imputed values in surveys with at most two missing item responses. The imputed values were equal to the adjacent values.[1] Our analysis also dropped individuals with inconsistent choices. A choice is classified as consistent if it allocates the same or a larger budget share to the future payment date than the previous choice with lower interest rate. As the interest rate increases, the law of demand dictates a weakly decreasing allocation to the earlier time point. To increase sample size, we manually changed the answers of surveys with at most two inconsistent choices. The new assigned values were equal to the adjacent values, and therefore the choices became consistent.[2]

50 percent of all complete and consistent responses had no variation. That is, parents

1. The imputation worked follows: (i) If the missing value was in a left-corner question (questions 1, 6, or 11), then we assigned the value of the right-adjacent question; (ii) Similarly, if the missing value was in a right-corner question (questions 5, 10, or 15), we assigned the value of the left-adjacent question; (iii) If the missing value is in an interior question (questions 2–4, 7–9, 12–14), the imputed value equals the adjacent questions only if they are the same. If the adjacent values are different, then no imputation is done.

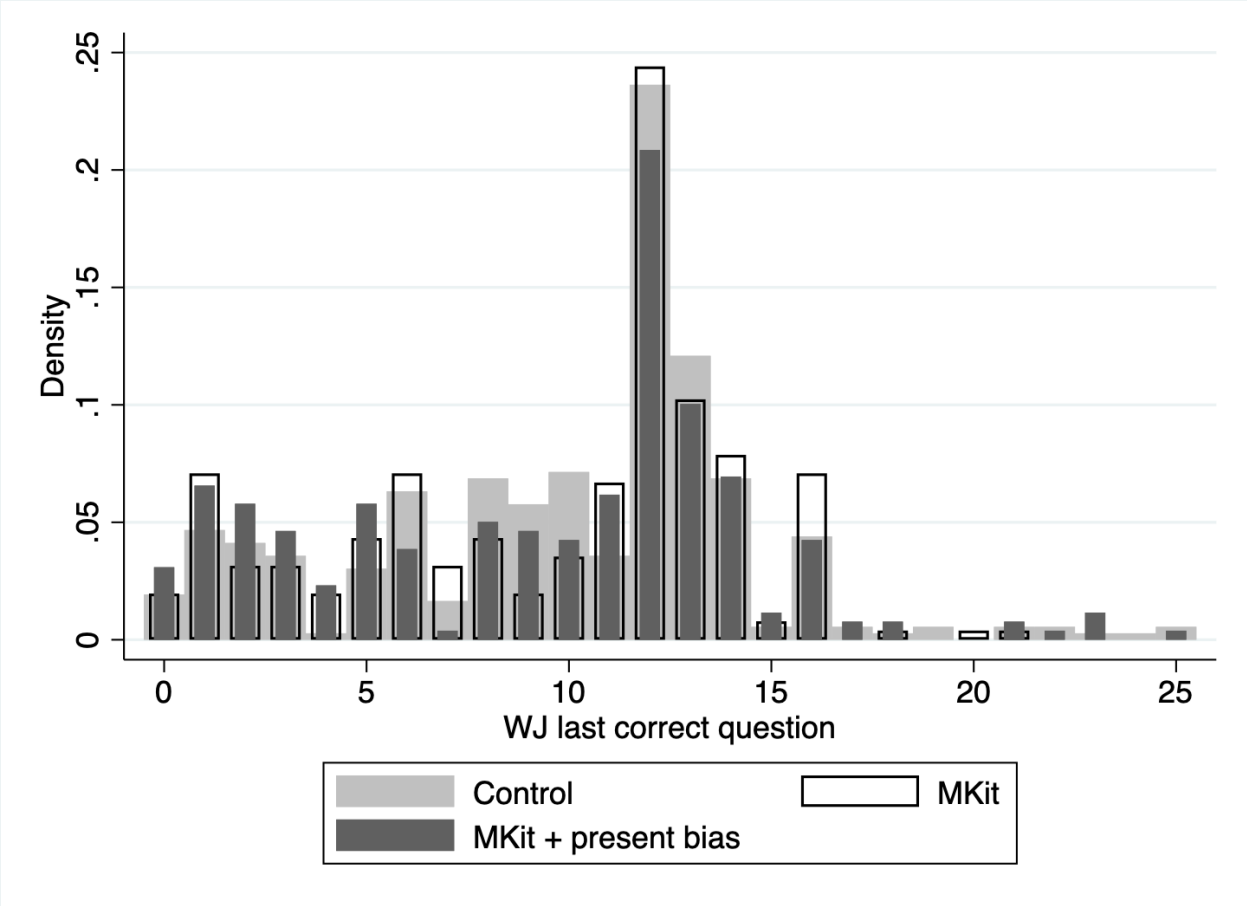
2. Surveys with two inconsistent choices that were contiguous (e.g. questions 2 and 3 or questions 2 and 4) were not modified. It was hard to know which adjacent value they should take.

chose the same answer throughout the task. Most of the surveys (68.2 percent) with no variation always chose the first option (i.e., parents preferred to receive the money early) while 23.3 percent always chose the last option (i.e., they preferred to receive $24 at a later date). The point estimates of the time-preference parameters cannot be recovered for these surveys with no variation. Parents who always preferred to take the money at an early date were classified as highly present oriented and with a high discount factor.

Table C.2.2 shows the estimates of time-preference parameters for parents that had variation in their responses.

C.2 Appendix Figures and Tables

Figure C.2.1: Distribution of the Last Correct Question in the Follow-up WJ by Treatment Group



Notes: This graph plots the distribution of the last question a child got correct in the WJ Applied Problems subtest. This test was administered in the first follow-up period.

Table C.2.1: Balance Test of Baseline Characteristics

	Control		MKit		MKit + present bias		Joint p-value	N
	Mean	S.D.						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Child characteristics								
Male	0.48	[0.50]	-0.04	(0.04)	0.01	(0.04)	0.50	1412
Age in months	46.87	[6.61]	0.18	(0.40)	0.61	(0.40)	0.28	1412
WJ z-score	-1.05	[0.93]	0.16**	(0.08)	-0.04	(0.08)	0.02	1412
Spanish consent form	0.40	[0.49]	0.00	(0.03)	-0.01	(0.03)	0.95	1412
Parental and family characteristics								
Parent's age	30.89	[6.71]	1.01	(0.69)	0.46	(0.67)	0.34	1098
Black	0.29	[0.46]	0.01	(0.03)	0.01	(0.03)	0.85	1098
Hispanic	0.64	[0.48]	-0.02	(0.03)	-0.01	(0.03)	0.82	1098
No HS diploma	0.23	[0.42]	0.07*	(0.04)	0.05	(0.04)	0.22	1098
HS diploma or GED	0.26	[0.44]	0.01	(0.05)	0.03	(0.04)	0.85	1098
Some college	0.36	[0.48]	-0.07	(0.05)	-0.05	(0.04)	0.32	1098
College or more	0.08	[0.27]	-0.04	(0.03)	-0.00	(0.02)	0.21	1098
Num. children	1.53	[0.78]	-0.00	(0.08)	-0.07	(0.08)	0.60	1198
Num. adults	2.22	[0.93]	-0.01	(0.08)	0.01	(0.08)	0.97	1198
Household size	5.23	[2.04]	0.11	(0.17)	0.01	(0.17)	0.79	1198
Both bio parents	0.49	[0.50]	-0.03	(0.04)	-0.03	(0.04)	0.73	1151
IEP indicator	0.08	[0.27]	0.03	(0.03)	0.04	(0.03)	0.34	1098
Birth weight $<$ 3.5 lb	0.02	[0.13]	0.02	(0.02)	0.02*	(0.01)	0.21	1098
Work full-time	0.23	[0.42]	0.02	(0.04)	-0.03	(0.04)	0.44	1098
Work part-time	0.28	[0.45]	-0.02	(0.05)	0.03	(0.04)	0.53	1098
In school	0.15	[0.36]	-0.01	(0.04)	-0.00	(0.04)	0.98	1098
Poor: below poverty line	0.71	[0.45]	0.00	(0.04)	0.04	(0.04)	0.54	1198
Dweck score	4.62	[0.81]	0.04	(0.09)	-0.05	(0.09)	0.65	1098
Math anxiety	2.31	[2.59]	-0.14	(0.26)	0.04	(0.25)	0.76	1098
Marginal effect 0 - 15	0.03	[0.11]	0.01	(0.01)	0.01	(0.01)	0.51	1098
Index negative barriers	0.09	[0.99]	-0.09	(0.10)	-0.09	(0.09)	0.53	1098
Index math activities	-0.13	[0.95]	0.17*	(0.09)	0.29***	(0.09)	0.01	1098
Index reading activities	-0.09	[0.96]	0.14	(0.10)	0.20**	(0.09)	0.10	1098
Impatience measure								
Impatient	0.73	[0.44]	-0.06	(0.07)	-0.06	(0.07)	0.57	606

Notes: Each row is a separate regressions where the dependent variable is indicated in the first column and explanatory variables are treatment group indicators and classroom fixed effects. Columns 1 and 2 show the control mean and standard deviation (in brackets). Columns 3 and 6 show estimated treatment coefficients and columns 4 and 6 report robust standard errors (in parentheses). Column 7 reports the p-value from a joint significant test of the MKit and MKit + present bias coefficients. Child variables come from the consent forms and child assessments, parent variables are obtained from the baseline parent surveys, and the present bias measure is from the money task. Column 8 reports the number of observations. The sample size varies for some parental variables because they were sometimes asked at the follow-up survey rather than at baseline. Asterisks denote significance level: * $p < .10$, ** $p < .05$, *** $p < .01$.

Table C.2.2: Estimates of Time-Preference Parameters Using the Money Task

Parameter	N	Median	Mean	S.D.	5th pctile.	95th pctile.	Min	Max
Baseline task								
Discount factor: δ	263	1.00	1.11	2.17	0.84	1.08	0.00	36.10
Present bias: β	263	1.00	1.61	2.54	0.00	11.69	0.00	11.69
CRRA curvature: α	263	0.99	0.77	0.51	-1.10	1.00	-1.10	1.00
Follow-up task								
Discount factor: δ	270	1.00	1.11	2.15	0.89	1.02	0.00	36.10
Present bias: β	270	1.00	1.05	0.61	0.00	3.00	0.00	3.00
CRRA curvature: α	270	1.00	0.84	0.31	0.00	1.00	0.00	1.00

Notes: This table shows estimated time-preference parameters. Estimation procedure is described in Appendix Section C.1. Values of β above the 95th percentile and values of α below the 5th percentile were truncated.