

Brigham Young University
BYU ScholarsArchive

Theses and Dissertations

2012-03-14

Propensity Score Methods as Alternatives to Value-Added Modeling for the Estimation of Teacher Contributions to Student Achievement

Kimberlee Kaye Davison
Brigham Young University - Provo

Follow this and additional works at: https://scholarsarchive.byu.edu/etd

 Part of the Educational Assessment, Evaluation, and Research Commons

BYU ScholarsArchive Citation

Davison, Kimberlee Kaye, "Propensity Score Methods as Alternatives to Value-Added Modeling for the Estimation of Teacher Contributions to Student Achievement" (2012). *Theses and Dissertations*. 3130.
https://scholarsarchive.byu.edu/etd/3130

This Dissertation is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.

Propensity Score Methods as Alternatives to Value-Added Modeling for the Estimation of Teacher Contributions to Student Achievement

Kimberlee Kaye Callister Davison

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Richard R Sudweeks, Chair
Joseph Olsen
K. Richard Young
Erika Feinauer
Lane Fischer

Department of Educational Inquiry, Measurement, and Evaluation

Brigham Young University

April 2012

Copyright © 2012 Kimberlee Kaye Callister Davison

All Rights Reserved

ABSTRACT

Propensity Score Methods as Alternatives to Value-Added Modeling
for the Estimation of Teacher Contributions to
Student Achievement

Kimberlee Kaye Callister Davison
Department of Educational Inquiry, Measurement, and Evaluation, BYU
Doctor of Philosophy

The purpose of this study was to examine the potential for using propensity score-based matching methods to estimate teacher contributions to student learning. Value-added models are increasingly used in teacher accountability systems in the United States in spite of ongoing qualms about the validity of teacher quality estimates resulting from those models. Using a large national dataset, teacher effects were estimated for 435 teachers using both value-added and propensity score-based approaches. The two approaches resulted in teacher effect estimates that were moderately correlated, with propensity score-based estimates more highly correlating with the value-added estimates as the matching ratio was increased. For many teachers' students, finding a set of matched control students was impossible unless the set of matching variables was reduced. Results suggest that many teachers have classroom compositions that are unusual, making evaluation of the teachers' impacts on student outcomes problematic. It was also found that, while value-added estimates were relatively insensitive to covariate inclusion choices or method of effect estimation, propensity score-based estimates were somewhat sensitive. Propensity score-based teacher effect estimates offer promise both for better accounting for classroom composition and student background variables and for indicating when a teacher's context is unique with respect to those variables, making the teacher's impact challenging to evaluate.

Keywords: value-added modeling, teacher accountability, teacher evaluation, propensity score analysis

ACKNOWLEDGMENTS

I would like to thank my BYU professors for instructing and motivating me and inspiring the ideas that led to this study. Thank you particularly to Dr. Sudweeks, an example of research integrity and lifelong curiosity. Thanks also to my parents, Douglas and Jan Callister, who taught me the importance of seeking excellence, and to Russell Everson, who always had enough time to listen.

I am especially grateful to my children for the real life lessons they taught me while I was pursuing my academic studies. To Michael, for demonstrating the importance of standing for my beliefs and having faith in my ideas. To Rachel, for showing me that I am always loved. To Joseph, who was continuously a support to me and an example to his siblings. To Daniel, for demonstrating how to set and work hard towards achieving ambitious goals. To James, who showed that we can create our own music as soon as we believe we can. To Anastasia, who proved that we can keep going and embrace life with zeal, no matter how hard our trials may be. And to Kristina and Max, who made it plain that learning is its own joy.

Lastly, I am grateful to the inspiring public school teachers I had as a child who taught me things that can never be measured by a test.

Table of Contents

Chapter 1: Introduction 1

Value-Added Modeling Assumptions 2

Propensity Score Analysis as an Alternative 3

Propensity Score Matching and Teacher Effects 4

Teacher Effects with Multiple Teachers 6

Research Questions 8

Chapter 2: Background and Literature Review 10

The Counterfactual and Random Assignment 10

Propensity Score Analysis 12

Teacher Accountability and Propensity Score Analysis 13

Cluster Sampling and Propensity Score Analysis 14

Multiple Treatment Effects and Propensity Score Analysis 18

Chapter 3: Method 21

Design 21

Sample 22

Measures 24

Procedures 25

Value-added models 27

Propensity score analyses 28

Generalized propensity score analyses 31

Comparisons 32

Chapter 4: Results 33
Sample Description 33
Covariate Selection 33
Value-Added Models 35
Sixty-six covariates 36
Seventeen covariates 37
Seven covariates 38
One covariate 38
Independent Propensity Score Models 42
Match quality 44
Teacher effect estimation 46
Sixty-six variables, all teachers 47
Seventeen variables, all teachers 57
Sixty-six variables, matchable teachers only 60
Generalized Propensity Score Analyses 61
Class-Level Matching 64
Chapter 5: Conclusions 65
Reflections on Findings 65
Treatment effects 65
Value-added models. 69
Propensity score analyses 71
Generalized propensity score analyses 73
Summary 75

Further Research 77

References 79

Appendix A: Comparability of Teacher Control Groups 87

Appendix B: Original List of 187 Potential Covariates 89

Appendix C: Reduced List of 111 Potential Covariates 95

Appendix D: Final List of 66 Covariates 98

Appendix E: Reduced Lists of 17 and 7 Covariates 101

Appendix F: Value-Added Covariate Parameter Estimates, Fixed Approach 102

Appendix G: Variables with Statistically Significant Imbalance after Matching 104

List of Tables

Table 1: Quintile Comparisons, Random and Fixed Effects with 66 Covariates 37

Table 2: Value-Added Effect Correlation Matrix using 1,7,17, and 66 Covariates 39

Table 3: Quintile Comparison Matrices, 66 and 1 Covariates 40

Table 4: Median Matched Distance, 66 Covariates, 199 Matchable Teachers 45

Table 5: Median Matched Distance, 17 Covariates, All Teachers 46

Table 6: Correlation Matrix for Matching and Effect Estimation Schemes when 66 Matching Variables and All Teachers were Used 48

Table 7: Concordance Correlation Coefficient Matrix for 66 Matching Variables and 435 Teachers 51

Table 8: Correlations of VA Fixed Effects (7 Covariates) with Propensity Score Effects (7 Covariates at Effect Estimation Stage) 52

Table 9: Matrix of Percent of 435 Teachers Falling within the Same Quintile across Effect Estimation Approaches and Matching Ratios Using 66 Matching Variables 52

Table 10: Percent of 435 Teachers Falling within the Same or Adjacent Quintile across Effect Estimation Approaches and Table 10: Matrix of Percent of Teachers Falling within the Same or Adjacent Quintile across Effect Estimation Approaches and Matching Ratios Using 66 Matching Variables 54

Table 11: Matrix of Percent of 435 Teachers Falling Moving from the Bottom 2 Quintiles to the Top 2 Quintiles or Vice Versa across Effect Estimation Approaches and Matching Ratios Using 66 Matching Variables 55

Table 12: Correlations of Propensity Score-Based Effects with 66 and 17 Matching Variables, All Teachers 58

Table 13: Correlation Matrix for Matching and Effect Estimation Schemes when 17 Matching Variables and All Teachers were Used 59

Table 14: Pearson Correlation Matrix for Matching and Effect Estimation Schemes when 66 Matching Variables and 199 Teachers were Used 60

Table 15: Correlation of PSA Effects with 66 or 17 Matching Variables, 199 Teachers 61

Table 16: Class Level Matched Ranking and Other Model Rankings 64

List of Figures

Figure 1: Concordance of fixed and random effects with 66 covariates 36

Figure 2: Concordance of fixed effects with 66 and 1 covariates..... 41

Figure 3: Concordance of random effects with 66 and 1 covariates..... 41

Figure 4: Diversion from line of perfect concordance of fixed VA estimates and propensity score estimates 56

Chapter 1: Introduction

United States policy-makers and legislators have become increasingly enamored with teacher accountability based on student outcomes (Eckert & Dabrowski, 2010; Goldhaber & Hansen, 2010; Newton et al., 2010; Sass, 2008). In 2010 alone, at least seven states passed legislation that encouraged the use of student outcomes for high stakes teacher evaluation purposes (National Conference of State Legislatures, 2010). State courts have a history of upholding the termination or non-renewal of contracts of teachers with low student test scores (*Massachusetts Federation of Teachers, AFT, AFL-CIO v. Board of Education*, 2002; *Scheelhaase v. Woodbury*, 1973; *St. Louis Teachers Union, Local 420, American Federation of Teachers, AFL-CIO v. Board of Education of the City of St. Louis*, 652 F. Supp. 425, 1987). In fact, in several recent Florida cases, school district decisions to terminate teachers were over-turned because the districts did *not* consider student test scores when making their decisions, even though the teachers were clearly incompetent by other standards (*Leon County School Board v. Waters*, 1986; *Sherrod v. Palm Beach County School Board*, 2006; *Young v. Palm Beach County School Board*, 968 So. 2d 38, 2006). In some cases, teachers have successfully fought back. For example, a Texas teacher who was fired for low student test scores was reinstated when those scores were shown to be the result of a poor school environment and not his fault (Lambert, 2008).

Many researchers have concerns about the use of student test scores to evaluate teachers, especially for high stakes purposes. McCaffrey, Sass, Lockwood, and Mihaly (2009) state, "For any performance-based personnel system to provide the correct incentives and enhance teacher quality, there must be a strong link between true performance and reward or retention" (p. 573). Value-added modeling is one attempt to improve this link between the estimation of a teacher's

impact on a student's test scores and true teacher performance by focusing on student gains in students' test scores, rather than status measures such as end-of-year achievement.

Value-Added Modeling Assumptions

One characteristic of value-added estimates of teacher quality, or *teacher effects*, is that they are relative—a teacher's effectiveness in improving student test scores is compared with other teachers in the same school, district, or state. No matter at which level(s) these comparisons are made, the lack of random assignment of students to classrooms can create imbalances in student potential to make gains across classrooms. These imbalances may lead to bias when value-added teacher effects are estimated, resulting in unfair teacher evaluations.

Value-added models generally attempt to account for these imbalances in the composition of classrooms by including some combination of student, family, classroom, and school covariates and student or school effects in the model, which is generally linear in form. To what degree any value-added model can really account for a teacher's classroom composition and estimate a teacher's contribution to student test scores without bias is one of the most commonly acknowledged issues in the value-added literature (Harris; 2009; Ishii & Rivkin, 2009; Koedel & Betts, 2009; Levine & Painter, 2007; Newton, Darling-Hammond, Haertel, & Thomas, 2010; Reardon & Raudenbush, 2009). One important concern is that other unmeasured covariates that are not included in the model may affect students' learning gains and may bias estimates of teacher effects (Baker et al., 2010; Harris; 2009; Rothstein, 2009). This requirement that all relevant covariates are included in the model in order to get unbiased effect estimates is referred to the *ignorable treatment assignment* assumption, and its violation is impossible to test directly. Only theory can suggest what covariates may be missing from a model (Rosenbaum & Rubin, 1983; Stuart, 2007; Thoemmes & Kim, 2011).

Regardless of whether all important covariates are included in the model, another assumption is necessary in order for estimates to be unbiased—the assumption of linearity in the relationship between the covariates and the dependent variable across all relevant values of the covariates. If the relationship is non-linear, bias may result, and this bias may be greater than if the covariates had not been included in the model at all (Cochran & Rubin, 1973). Even if a linear model appears to have good fit, a linear model is fitted to the values of the covariates under the control condition, and so extrapolation is involved when assuming the relationship of the covariates follows the same linear trend in the treatment group (Stuart, 2007). Nonlinearity in value-added teacher effect linear regression models used to estimate teachers' value-added effects has been shown to be a concern (Reardon & Raudenbush, 2009).

Rubin (2001) demonstrated that regression analysis for the comparison of a treatment and control group in an observational study should only be trusted if specific conditions regarding covariate balance in the treatment and control groups are met. Essentially, covariate-adjustment in a regression model is only reliable if the treatment and control groups have similar distributions on those covariates, and even minor differences in the distributions can cause problems (Rubin, 2001; Stuart, 2007). However, when teacher assignment is modeled as a treatment in a value-added regression model, covariates are often added to the model precisely *because* the distribution of covariates is expected to be non-equivalent from classroom to classroom. To the degree that the covariates are really distributed differently across classrooms, bias in the value-added effects is likely.

Propensity Score Analysis as an Alternative

As an alternative, propensity score analysis (PSA) is a family of statistical techniques which approximates random assignment by matching each treated subject to the closest possible

control in the larger untreated sample (in terms of propensity to be chosen for the treatment), thereby allowing estimation of unbiased treatment effects in the absence of random assignment (Austin, 2011; Dattalo, 2010; Guo & Fraser, 2010; Rosenbaum & Rubin, 1983). Instead of (or in addition to) controlling for covariates at the treatment effect estimation stage, propensity score-based methodologies use these variables for identifying the best possible control group for comparison with the treatment group. Rosenbaum and Rubin refer to this process as *balancing* on the covariates, and they show that this balancing makes possible an unbiased estimate of the *average treatment effect* (ATE) for the population, provided treatment assignment is strongly ignorable. In other words, the estimate of the ATE will be unbiased if and only if the correct covariates are included in the model. After matching, the treatment effect may be estimated using a linear model, controlling for the covariates, but the treatment and control groups now have similar covariate distributions and so the assumptions of linear modeling are better met.

Ignorable treatment assignment is an assumption-in-common for bias-free effects resulting from both regression and PSA models. However, the other significant regression assumption related to bias, linearity, is *not* a requirement of PSA models (Guo & Fraser, 2010; Rosenbaum & Rubin, 1983; Rubin, 2005; Thoemmes & Kim, 2011). This suggests that, assuming identical covariates are included, PSA estimates are at less risk of bias than are regression estimates.

Propensity Score Matching and Teacher Effects

Applying PSA methods to the estimation of one teacher's effectiveness as compared with some control group is straightforward on the surface. If assignment to a specific teacher is equated with the treatment state, and non-assignment to that teacher is equated with the control state, then PSA methodology might be an alternative to value-added modeling for estimating the

teacher's effect on student achievement, balancing the distributions of covariates in the treatment and control states and avoiding the need for an assumption of linearity in the relationships between the covariates and student achievement. Essentially, one teacher's students would be compared with the best possible control group of students from the entire school, district, state, or nation. A teacher effect would be estimated by comparing the teacher's students with the selected control group as done in a random experiment. Any given teacher's PSA-based effect estimate should have less bias than the same teacher's effect estimated from a value-added model, using the same covariates, if non-linearities are present in the relationship between covariates and student achievement or if the distribution of covariates differs across teachers.

The term *propensity score* typically refers to the probability of treatment assignment, conditional on a set of covariates. Rosenbaum and Rubin (1983) describe the propensity score, e(x), as a type of *balancing* score, which reduces a multidimensional set of covariates to a single dimension. If teacher assignment represents treatment, then it is awkward to think of e(x) as the probability of teacher assignment, given a set of covariates. Obviously, a student in Ohio has no probability of assignment to a teacher in Tennessee. We might instead think of e(x) as a function of the covariates that *predicts* whether a student is a member of teacher *j*'s class. Another point of view is that the propensity score is a function that identifies covariates that are not balanced across classrooms. If e(x) is exactly equal for two students, one in the teacher's class and another outside the teacher's class, then we cannot predict, based on the covariates, which student is most likely to be a member of teacher *j*'s class. While the two students may have different values on the covariates, each vector of covariates is just as likely to be in teacher *j*'s class as is the other (Joffe & Rosenbaum, 1999). Assuming every variable which causes teacher *j*'s students to achieve high test scores, other than the class membership itself, is included in the

model, then the only remaining possibility is that the test scores were a result of being in teacher *j*'s class. In other words, if no confounding variables are missed, and if pairs or sets of students are found with perfectly matching propensity scores, then any differences in the student achievement within versus without the teacher's classroom must be related to belonging to that classroom.

While Rosenbaum and Rubin (1983) show that certain effect estimates will be unbiased using propensity score matching, they do not, of course, suggest that factors that are confounded with treatment status can be separated. For example, if teacher *j*'s students have unusually high access to materials, specialist support, or parent volunteer hours, or if peer-interactions affect students in the class, then the *teacher effect* will no more represent quality of the teacher than in value-added models. If school-level variables are not accounted for, then the teacher effect also contains inputs to student achievement due to schools. A more appropriate term that encompasses these other impacts on student achievement would be *classroom effect* (Rothstein, 2009), but *teacher effect* will be used here as is most often done in the literature, with the understanding that other factors than the teacher may impact that effect.

Teacher Effects with Multiple Teachers

Complications, however, arise in the PSA procedure because we are generally interested not only in a particular teacher's effectiveness, but in that teacher's effectiveness relative to other teachers. Even if all we care about is the *absolute* effectiveness of individual teachers, a comparison of PSA results to VA results is impossible if the entire set of PSA teacher effects cannot be ranked.

The problem is rooted in the structure of the sample. Essentially, a large sample of N students is clustered into k groups, or assigned to k teachers/classrooms/treatments.

Teacher/classroom j receives n_j students, and the remaining N- n_j students represent the control group for teacher j. Teacher/classroom m receives n_m students, none of whom can belong to teacher j's class, and the remaining N - n_m students (which include teacher j's students) belong to teacher m's control group. The complications resulting from applying propensity score matching to this sample structure include the following:

1. No two teachers' effect estimates will be *exactly* comparable because no two teachers' control conditions will be defined exactly the same way (Control condition for teacher j = not taught by teacher j; Control condition for teacher m = not taught by teacher m). Appendix A demonstrates that this lack of overlap becomes trivial in the estimation of the average treatment effect (ATE) if k, the number of teachers, is large.
2. Students in any given teacher's control group are clustered to schools before assignment to classes, and to teachers after, and so both assignment probabilities and responses may be correlated.
3. Classroom (teacher) membership is mutually exclusive across all classrooms. Estimating the probability of belonging to a particular class, given the covariates, should ideally take into account this exclusivity.

One alternative is to use Imben's (2000) *generalized* propensity score which is used to estimate unbiased treatments effects when there are more than two treatment groups. This approach uses a multinomial logit function in place of the logit function and results in a vector of propensity scores instead of a single, scalar, estimate. The limitation of this method is that, as the number of treatment groups increases, estimation of the multinomial logit function becomes problematic at a rate that is much more than linear, but instead related to $\binom{k}{2}$ if k represents the number of treatment groups. Convergence problems are common, even when k is relatively

small. A possibility is to use the method iteratively for *sets* of teachers, with teachers deemed to be most similar grouped to the same set. An advantage of this approach, in addition to accounting for exclusivity and comparability (at least within sets), is that the grouping of students to teachers is accounted for—there is no unaddressed clustering.

However, the counterfactual is the same whether a series of independent logistic regressions are used to estimate propensity scores or one multinomial logistic regression model is run, provided the same classes of students and variables are used in both analyses. In fact, the multinomial approach estimates the same equations as in the independent approach. The only practical difference between the approaches is that the equations are estimated simultaneously rather than independently. This difference affects the generalized propensity score estimation process two ways. First, the generalized propensity scores across groups for an individual are constrained to sum to 1.0 across groups when multinomial models are estimated. Second, the multinomial regression models are harder to estimate (Kleinbaum & Klein, 2002).

Research Questions

Using annual kindergarten mathematics item response theory (IRT) ability gains from a large national dataset with a rich set of covariates, teacher effects for one teacher per school (due to data limitations) were estimated and the following questions were addressed:

1. How do estimated teacher effects with identical sets of covariates compare using the following methodological approaches:
    a. Covariate-adjusted value-added models with random effects
    b. Covariate-adjusted value-added models with fixed effects
    c. Propensity score-based matching, with teacher propensity scores estimated independently using logit regression.

   d. Generalized propensity score-based estimation, with propensity scores estimated jointly within specified strata using multinomial logit regression.

2. What methodological problems and limitations are discovered using propensity scores and generalized propensity score approaches?
3. For pairs of classrooms with similar covariate distributions, to what extent do each of the four methodologies rank the teachers the same way?

The study was limited to comparisons of teachers across (rather than within) schools. The purpose was to test methodology rather than to make actual conclusions about the specific teachers represented in the dataset. In addition, conclusions may have been sensitive to the particular nature of the dataset, sample, and variables measured, and so may not be directly generalizable to datasets with a different structure. Effect estimates may have been, in each approach, biased due to any violations of the ignorable treatment assignment assumption (unmeasured covariates). In addition, school and teacher effects were confounded because the sampling design used in the national dataset makes analysis of more than one teacher effect per school problematic (too many classrooms with extremely small sample sizes).

This study is particularly significant in the light of an increased national emphasis on evaluating teachers through student outcomes and a growing tendency for states to create legislation and policy that encourages the use of student test scores to evaluate teachers. If high stakes decisions about teachers are made based on student test scores, it is critical that estimates of teacher quality are as well-founded as possible, and that we know when they are not.

Chapter 2: Background and Literature Review

Prior literature was evaluated to include statistical theoretical foundations related to causality. In addition, propensity score literature was evaluated for connections to teacher evaluation and the sampling structure used in the study.

The Counterfactual and Random Assignment

The Neyman-Rubin counterfactual framework is based on the theory that two possible outcomes exist for every unit in the population of interest—the outcome with treatment and the outcome without treatment (Guo & Fraser, 2010; Morgan, 2001; Morgan & Winship, 2007; Rosenbaum & Rubin, 1983; Rubin, 1974). In applying the framework to achievement-based teacher evaluation, student *i* would have a potential end of year test score r_{1ij} with assigned teacher *j*, and a second potential test score r_{0ij} if not assigned to teacher *j* (assigned to a theoretical *average* teacher). All other factors being held constant, comparison of the outcomes, $r_{1ij} - r_{0ij}$, would be the effect of assigning teacher *j* on student *i*'s test score, and $E(r_{1j}) - E(r_{0j})$ would represent the true average treatment effect (ATE) for teacher *j* for all students in the population of study, or the *teacher effect* (notation adapted from Rosenbaum & Rubin, 1983).

The problem is that r_{0ij} is *counterfactual*, a potential outcome that is never observed, for the students in the teacher's class, and r_{1ij} is counterfactual for the students *not* in the teacher's class. The Neyman-Rubin framework estimates the average treatment effect of teacher *j* by subtracting the mean outcome of the untreated (students not assigned to teacher *j*) from the mean outcome of the treated, the students assigned to teacher *j* (Guo & Fraser, 2010; Rosenbaum & Rubin, 1983).

This treatment effect is based on several assumptions. First, the ignorable treatment assignment assumption requires that, for an effect estimate to be unbiased, assignment to

treatment must be independent of either potential outcome, after conditioning on covariates. Additionally, for the ATE, the probability of assignment to treatment for all values of a covariate must be greater than zero and less than one (Guo & Fraser, 2010; Rosenbaum & Rubin, 1983). When estimating teacher effects on achievement, student *i*'s assignment to teacher *j* must be unrelated to r_{0ij} and r_{1ij}, the student's potential achievement, after controlling for all covariates in the model. However, this assumption is generally violated in educational settings. Student assignments to teachers are often correlated with student achievement potential in ways that cannot be easily measured.

The ignorable treatment assignment assumption, also referred to as *exogeneity* or independence of the error term and the independent variable (e.g., teacher assignment), is a particularly important assumption in regression models such as value-added models. Violations of the assumption lead to biased and inconsistent estimates of treatment effects (Guo & Fraser; Rosenbaum & Rubin). Essentially, estimates of teacher effects on student test score gains may be over or understated if assignment to teachers is not independent of student potential to make gains. A large body of research suggests that non-ignorable treatment bias results in biased value-added estimates of teacher quality. The potential bias resulting from the non-random assignment of students to teachers is the most consistently reported concern of both value-added researchers and skeptics.

Many approaches have been offered in the attempt to correct for overt treatment bias (bias based on measurable variables) in observational studies. However, in no case can hidden bias (bias due to *unobserved* variables) be corrected for, suggesting that the random experiment remains the "gold standard" (Guo & Fraser, 2010, p. 38). Guo and Fraser discuss four approaches to correction for treatment bias. These include (a) Heckman's sample selection

model, (b) propensity score models, (c) matching estimators, and (d) nonparametric propensity scores. Other methods include (a) regression estimators, (b) combined approaches, (c) Bayesian approaches (Imbens, 2004), as well as (d) regression discontinuity, (e) instrumental variables, (f) interrupted time series, (g) differential growth models, and (h) analysis of covariance (Winship & Morgan, 1999).

Propensity Score Analysis

A *propensity score* is the probability of assignment to treatment, given a set of covariates. Essentially, the propensity score reduces a multi-dimensional vector of covariates to one dimension, and a treated unit is matched to a control unit (or several) with a similar estimated propensity score. Treatment and control units are not explicitly matched based on the values of each of the covariates. However, Rosenbaum and Rubin (1983) show that the two groups do end up balanced overall on the covariates. Two units with the same propensity score may differ with regard to any one covariate, but that difference will be due to chance rather than systematic, and the two units will have the same chance of being assigned to treatment, given the propensity score, as in a random experiment (Guo & Fraser). This means that, given the true propensity score and no missing covariates, the treatment effect will be unbiased.

Given *exact* matching on the *true* propensity score, Rosenbaum and Rubin (1983) prove that the mean of matched paired differences is an unbiased estimate of the true average treatment effect (ATE). The probability distributions of covariates in the treatment and control groups are balanced, matching the probability distributions of either matching exactly on all covariates or random assignment. In addition, they show that matching exactly on the *estimated* propensity score results in sample balance if all relevant covariates are included in the model and the model is of correct form. When post-matching analysis is done, the authors indicate that effect

estimates based on matched samples are more robust to departures from true model form than are even models based on random samples, because extrapolation on values of the covariates is less likely to happen. This lack of extrapolation is dependent on a large region of common support for the treated units and potential controls—the assumption that there is sufficient overlap between the covariate distributions of the treated and controls. PSA-based estimates also have lower standard errors than random assignment estimates because the distributions of the covariates are closer than happens by chance. Rosenbaum and Rubin also suggest that matching methods are particularly valuable in studies that have small treatment groups, large potential control groups, and a large number of covariates—as can be true in studies which compare teachers. The small treatment group sample size (degrees of freedom limitations) prevents the inclusion of so many covariates using other methods. When matches are not exact (propensity scores in the treatment and control groups are only very close), simulations show that treatment effect bias is still greatly reduced, compared with unmatched comparisons.

However, propensity scores may be inaccurate if the model predicting the scores is not correct. As a result, researchers check to see whether the propensity score-based matching of treatment and control groups has sufficiently balanced the data on each of the covariates, as it would if the model were correctly specified. If the two groups are not balanced, then likely the form of the variables is not correct—higher order or interaction terms may be needed (Guo & Fraser, 2010). In addition, estimated propensity scores may be biased if any necessary covariates are not included.

Teacher Accountability and Propensity Score Analysis

A search of literature regarding the use of propensity score analysis for the purpose of teacher evaluation for accountability purposes uncovered no articles. ERIC, EconLit, and

Econpapers (economics working papers) were searched using terms such as *propensity score and ("teacher effect" or "school effect" or accountability)* or *"value-added" and "propensity score"*. While research was discovered that used propensity score analysis for evaluation of educational programs, interventions, or other malleable factors, none appeared to address accountability at either the school or teacher levels. Hahs-Vaughn and Onwuegbuzie (2006) state that propensity score methods have been slow to take hold in the field of education and the social sciences.

Cluster Sampling and Propensity Score Analysis

When teacher effects are estimated, clustering of students generally needs to be accounted for at two levels—the class level and the school level. In addition, clustering needs to be accounted for at two time points—before and after assignment to classes. Before matching, when propensity scores are estimated, we are only interested in covariates which affect students *before* assignment to treatments (Rosenbaum & Rubin, 1983). Only school-level covariates, or possibly previous year class-level covariates, should cause within-group correlation in student ability before assignment. In this study, however, only one class from each school was represented in the study. As a result, school and class levels were confounded. Ideally, this sample design would be accounted for when propensity scores were estimated as part of the logit regression.

In addition, *after* assignment to classes student responses were expected to be correlated at both the class and school levels. This means that both school- and class-level nesting would ideally be accounted for at the effect estimation stage for each teacher's potential control group. The problem in this study, however, was one of separability. Teacher assignment was confounded with both clustering and classroom or school covariates. At the logit-model

estimation stage, the higher-level covariate or effect would perfectly predict the dependent variable (teacher assignment), and at the effect-estimation stage, the higher-level covariate or effect would be collinear with the treatment assignment covariate. A review of the literature reveals no method of accounting for clustering in a propensity analysis other than to adjust the logit and effect models as described.

Literature on cluster sampling in propensity score analyses was searched, and the goal was to find methodological literature addressing the use of propensity score analysis when the sample included clustering of any kind. In addition, studies were examined that attempted to apply propensity score analysis to clustered samples. Articles that addressed cluster methods without relating them to PSA or which involved clustered samples but did not directly address the clustering were excluded. Searches were conducted in ERIC, EconLit, and Econpapers, using a variety of combinations of terms such as: *multilevel*, *hierarchical*, *cluster*, *propensity score*, *matching estimators*, and *complex samples*. References in the articles that were discovered were gleaned for missed literature.

Thirteen peer-reviewed articles or working papers were found which at least minimally addressed the issue. A review of the literature suggested that none addressed sampling structures identical to the one used in the current study.

Thoemmes and West (2011) appeared to summarize the existing methodological literature and proposed several models based on two of the most common clustering structures---treatments assigned within clusters and clustering which is *incidental* to treatment status. For the first case, which Thommes and West called the *narrow inference space*, treatments are assigned in such a way that every cluster contains both treated and control units. The authors suggested that the logit function used to estimate propensity scores be modeled as a multivariate linear

model with both random and fixed effects. A hierarchical or multilevel linear model would be fit, and the clustering would be accounted for by random intercepts and random slopes, if desired. The drawback of this model, the authors explained, is that propensity scores estimated within each cluster would not be comparable to propensity scores from other clusters. As a result, matching could only be done within-clusters. This narrow inference space model did not match the current study. In the current study, treatments were equivalent to teacher assignments—it was impossible to assign multiple teachers/treatments within clusters.

Thoemmes and West's (2011) second scenario, the *broad inference space*, is a sampling design in which units are assigned to treatment or control status without regard to clustering, but clustering is present. For example, participants may be assigned to one of two groups without regard to their school membership, but some will happen to belong to one school and some to another—and both treatment conditions are not necessarily present within each cluster. In this case, the authors suggested using the same hierarchical logit model, but dropping all random effects. Level-2 parameters would not be estimated, but level-2 covariates would be included in order to balance the samples with the clustering taken into account. Propensity scores would be comparable across clusters and matching could be done at the across-cluster level. Griswold and Localio (2010) described the possible approaches as *ignoring* clustering, *no pooling* (Thoemmes and West's narrow inference model), or *partial pooling* (the broad inference model).

The broad inference case more closely matched the present study because controls for each teacher's class were sampled in such a way that clustering was incidental. However, in this study, the treatment group represented one intact cluster and so adding level-2 covariates to the logit model would result in perfect prediction of the dependent variable, as described previously.

With either model, Thoemmes and West (2011) suggested that post-matching analysis should involve using a hierarchical linear model with either fixed or random effects or other methods that correct standard errors to account for clustering. This approach would have created collinearity problems in the current study due to the school-teacher level confounding.

Stuart (2007) discussed using propensity score matching to compare schools when the data consist of school-level data such as school-wide means. She suggested that when a treatment is assigned to an entire group, such as a school, the group be considered the unit of study. This method, as applied to the current study, would mean that matching would take place at the teacher/classroom level instead of the student level, and valuable information about the distribution of students within the classroom would be lost.

Hong (2010) suggested an alternative but similar method to propensity score estimation, *marginal mean weighting through stratification* to adjust for selection bias in multilevel settings. The remaining researchers in this pool of literature estimated multilevel models at either the logit regression (propensity score estimation) stage by using either of Thoemmes and West's (2011) approaches, or at the effect estimation stage, or both (Arpino & Mealli, 2008; Guo & Zhao, 2000; Hong & Raudenbush, 2005; Hong & Yu, 2007, 2008; Kim & Seltzer, 2010; Mulrow, 2010; Schreyogg, Stargardt, & Tiemann, 2011; Smyth, 2008). None of these papers involved any sampling design other than assignment to treatment within clusters or incidental clustering to the treated and untreated units, and none solved the problem created by the sampling design in this study.

The primary reason to account for clustering, however, is to avoid inflated standard errors. When clustering is ignored, estimates are unbiased unless theory suggests random slopes (interactions between the treatment assignment mechanism and the cluster), should be modeled.

In the context of the present study, at both stages of modeling (propensity score estimation and effect estimation), bias was important and standard errors were not. Propensity-score estimates were used for matching without considering standard errors of the estimated regression coefficients. Moreover, value-added effect estimates are typically used by researchers for ranking teachers without considering standard errors, and the same practice was followed with the propensity score-based estimates in this study. Rankings of teachers across models was compared without considering statistical significance of the estimates.

Multiple Treatment Effects and Propensity Score Analysis

Attempts have been made in prior research to apply propensity score estimation procedures when there are more than two treatment groups. Joffe and Rosenbaum (1999) proposed extending propensity score estimation methodology for cases in which levels of the treatment variable are *ordinal* by using a McCullagh's ordinal logit model, and there is a pool of literature built on that approach (Lu, Zanutto, & Hornik, 2001; Zanutto, Lu, & Hornik, 2005).

Current approaches for nominal-level treatment groups appear to stem from the work of Lechner (1999) and Imbens (2000), who cross-cite each other. Imbens (2000) proposed a model in which the multiple (3+) levels of the treatment variable could be nominal by defining a *generalized propensity score*, $r(t,x)$, for a unit as the unit's probability of assignment, given the covariates. As in Lechner's method, the *score* is a vector:

$$r(t,x) = \mathrm{P}(T = t \mid X = x) \qquad (1)$$

In the case of nominal levels of treatment, the multinomial or nested logit is used to estimate the generalized propensity score. One limitation of this methodology is that, unlike with Rosenbaum and Rubin's propensity score, causal effects cannot be found for *specific values* of $r(t,x)$, or for strata based on those values. The function of covariates used in calculating

generalized propensity scores differs across treatment groups. As a result, subjects in different treatment groups cannot be matched based on their propensity scores (see also Lu, Zanutto, & Hornik , 2001). However, Imbens demonstrated that the average treatment effect across strata has a causal interpretation. This is found by two steps. First, $\beta(t,r)$, the conditional expectation of the response variable given both the treatment level and the propensity score, is estimated. Essentially this is the vector of predicted values found when the response is regressed on t and $r(t,x)$. Next, for each level of treatment, this vector of predicted values is averaged over the distribution of the covariates. This gives us the estimated expected value of the response, given the covariates, at each level of treatment. These expected values for each level of treatment should be comparable. In Imbens' context (medical treatments), differences in the expected values are assumed to be causal results of the treatment choice. Imbens' method appears to be used often in the literature (Ertefaie & Stephens, 2010; Foster, 2003; Gingerich, 2010), and extensions to the case of continuous treatment variables have been made (Doyle, 2011; Fryges, 2009). As described in Chapter 1, the feasible application of Imben's approach to the current study is limited by the large number of treatment levels when teacher effects are being estimated.

While Lechner's theoretical approach was different than Imben's, it was similar in that a multinomial logit was used to estimate a propensity vector. He suggested that the propensity-score advantage of reducing a set of covariates to a single dimension only remains an advantage with the generalized propensity score if the number of treatment levels is less than the number of covariates (the propensity score vector is low-dimensional). If treatment effects of a large set of treatment effects are being estimated, then he suggested that his method poses no advantage over directly matching on the covariates, in terms of simplicity and dimension-reduction. However,

in the current study, matching directly on the covariates did not account for nesting as does using a generalized propensity score approach, and so there may still have been advantages to using a multinomial logit function to estimate a propensity score vector, especially within groups of similar teachers.

Imai and van Dyk (2004) expanded on Imbens (2000) by a further generalization of the propensity score—the *propensity function*, which allows the treatment variable to be nominal, ordinal, continuous, or even multivariate. For a nominal treatment variable, however, the propensity function is essentially the same model proposed by Imbens, with the same limitation that it will have as many dimensions as there are treatment levels.

The approaches suggested by these authors did have the advantage over the scalar propensity-control method in the context of this study, however, in that clustering was not an issue. Using their approaches, a cluster would always be perfectly confounded with a teacher-treatment group.

Chapter 3: Method

The study was a comparison of statistical methodologies using archival data. These methodologies included both common and innovative strategies for teacher effect estimation.

Design

Four different approaches to estimating the same set of teacher effects were used: (a) value-added modeling with random effects, (b) value-added modeling with fixed effects, (c) scalar propensity score matching, and (d) the generalized propensity score approach.

Existing data from a national dataset were used for this study, as it met the requirement of containing a large number of variables to be used in the propensity score analysis. This minimized the risk of violating the ignorable treatment assignment assumption. The Early Child Longitudinal Study (ECLS-K) was a national longitudinal study that followed one cohort of students from kindergarten through grade 8, beginning in the fall of 1998. A purpose was to investigate the impact of background variables on educational abilities, outcomes, and gains for young elementary students, which made the dataset ideal for this study. Data were collected at the beginning and end of kindergarten and grade 1 and at the ends of grades 3, 5, and 8. Students, teachers, administrators, and parents were interviewed or surveyed, and students were given a variety of assessments (Tourangeau et al., 2009). This dataset was unusual among national datasets in that assessments were initially repeated at a one-year interval, which was necessary for the calculation of teacher effects in this study. The ECLS-K study was a repeated measures design using the same students from year to year, with some attrition.

In the first year of the ECLS-K, children were chosen using a multistage probability sampling design. At the first stage, 100 counties, or in some cases groups of counties, were chosen with probability proportional to the number of five-year-olds, with oversampling of

certain ethnic groups and some stratification based on a variety of factors such as metropolitan size, percent minority, and per capita income. At the second stage, a total of 1413 public and private schools with kindergartens were chosen. Of these, 136 were found not to have kindergartens after sampling, and some declined to participate initially, leaving a final fall kindergarten sample of 1018 schools. Within each school, kindergarteners were randomly selected for inclusion in the study, with certain ethnic groups oversampled. The target number of students from each school was 24, but some schools, presumably smaller schools or those with insufficient numbers of students in the oversampled ethnic groups, had fewer students included. In the end, 19,985 students were included in the final fall kindergarten sample (Tourangeau et al., 2009). County-level information appeared to have been suppressed in the public-use dataset, but region, location type (urbanicity), and (coded) school assignments were available.

Sample

Because the non-public schools included in the sample tended to more often have small numbers of students per school sampled, only public schools were included in the present study. This reduced the potential sample to 628 schools and 14,017 students. From this sample, 2176 students who switched teachers or schools during the kindergarten year or who joined the study after the fall data collection were eliminated, leaving a potential sample of 11,451 students.

In this remaining group, schools had a mean of 19.9 students included in the study with a standard deviation of 3.7. Only thirteen schools had nine or fewer students in the study, but the students in the schools may have been assigned to any feasible number of teachers and so many class sizes were very small. From each school, the teacher was chosen with the largest number of participating students. If a school had no classroom containing at least five participating students, no teacher effect was estimated for that school. This selection criterion was used

because it gave an almost-representative sample of schools in the dataset, without over-inclusion of small sample sizes at the classroom level. The resulting sample for which teacher effects were estimated included 435 teachers (and, thus, 435 schools) and 4617 students. However, all 11,451 public school students who remained with the same teacher for both the fall and the spring data collections were used as potential controls for each of the 435 classes.

For the value-added and scalar (logit regression) propensity score approaches, this entire sample was included directly. For the generalized (multinomial logit regression) propensity score approach, these 435 teachers were classified into strata. A logical stratifying variable was suggested by the first stage of the ECLS-K multistage sampling design—the county. Classroom assignments across counties were nearly independent, as they would be across states. However, the county membership was not included in the public-use dataset and inquiry suggested the variable was not available through the restricted-use dataset. As a result, an alternative stratification scheme was necessary.

The public use dataset offered only two other possible stratification variables—region (Northeast, Midwest, South, and West), and location type (central city, urban fringe and large town, rural and small town). It seemed reasonable to assume teacher assignments were similar within each of these geographically-related stratum, making 12 strata possible. However, the large sample of teachers meant the stratum each included a large number of treatment groups, or classes, ranging from 20 to 50, depending on the stratum. A multinomial logit regression was unable to handle this many treatment groups. A reasonable solution was to randomly create sub-strata within each stratum. The size of the sub-strata were chosen such that they were as large as possible, while still allowing the multinomial logit regressions to converge using the desired covariates. The maximum number of teachers which could be chosen from each sub-stratum

depended on the actual covariates used in the model and any estimation problems encountered when fitting the multinomial regressions. This maximum was determined in this study by trial and error.

Measures

The response variable was ability estimates from the mathematics test given to kindergarten students at the end of the school year (C2R4MSCL). These 3-PL IRT-based estimates had been rescaled to indicate the total number of items a student with that ability was expected to have answered correctly. Scores from the same test given at the beginning of the school year (C1R4MSCL) were used as a covariate. When the tests were administered, a routing form was used to determine whether the child fell into the high, medium, or low group on the mathematics test. Next, the assessment was administered individually by a trained proctor who presented questions to the child and entered responses into a computer (Tourangeau et al., 2009). The reliability of the kindergarten mathematics test scores was .92 in the fall and .94 in the spring. The mathematics test was based on the 1996 NAEP framework, and was vertically scaled (Rock & Pollock, 2002). The test was proprietary and item level data were not available in the public use dataset.

Other student, family, and school background variables from the dataset were used as covariates in the value-added models and as matching variables in the propensity score analyses. The ECLS-K dataset encompasses hundreds of variables, ranging from measurements or background information taken on individual students to information gleaned from parent, teacher, or school surveys. While propensity score analyses do not limit the number of variables which can be used for matching, data used for real life teacher evaluation does not generally include a large number of variables, and regression-based value-added estimates have limitations

on both the number of and correlations between the covariates chosen. The important question of *which* variables are the best combination of practically measureable and meaningful for use as covariates in ongoing teacher assessment was not investigated in this study. Variables were chosen for this study as described below.

Procedures

Guo and Fraser (2010) suggested a procedure for selecting matching variables (covariates) when doing standard propensity score analyses: First, check the treatment group and the potential control group as a whole for balance on each potential matching variable, using appropriate statistical tests. Include in the model as a matching variable any with significant imbalances between the two groups; second, after matching, conduct the same statistical tests for balance on the treatment and smaller selected control group; third, if any significant differences in balance remain on any variable, reconfigure the matching variables to include higher order terms or interactions for that variable and retest.

One complication with this study was that 435 propensity score analyses (12 or more generalized PSA's) were performed, one for each teacher or group of teachers. Following Guo and Fraser's (2010) covariate selection procedure may have created different model formulations for each of the 435 teachers (or 12 or more generalized PSA strata), meaning different matching variables would have been selected. In order to create uniformity in the matching variable choice, it was necessary either to increase the number of variables used in all analyses (including interactions or higher order terms) so that each of the PSAs or generalized PSAs resulted in balance, or to tolerate greater imbalances in some of the treatment-control group pairs for the sake of parsimony. While parsimony is not generally a goal of PSA or generalized PSA, having fewer matching variables was desirable for both the VA analyses and for the sake of practical

applicability of the methodology to teacher evaluation. It also became evident that too many matching variables created poor matching in some of the PSA analyses. In addition, multicollinearity among any chosen covariates was a concern in the VA analyses.

Approximately 200 potential student and parent-level covariates were chosen from the ECLS-K dataset which were subjectively determined to best meet the criteria of potentially unbalanced across classrooms and potentially related to kindergarten mathematics score gains. From these potential covariates, any with significant missing data or minimal variability in the responses were removed from consideration, including categorical variables with the majority of the responses at one level. Those remaining were assessed for multicollinearity and removed, if necessary.

For the remaining variables, Guo and Fraser's (2010) first step, the check for balance on each treatment group (teacher's class) versus all potential controls, was adapted. Because the goal was to identify covariates which tend to be distributed non-uniformly across teachers, a one-way ANOVA or chi-square was estimated for each potential covariate. Variables with the lowest p-values (least similar covariate distribution across teacher classrooms) were chosen for consideration in the study. The beginning of year (fall) kindergarten general mathematics test IRT ability estimate, which was used as a covariate or matching variable, was used in addition to the other selected covariates. After the measures to be used in the study were chosen, values for missing data on the covariates were imputed using maximum likelihood estimation.

Because the multinomial models were not estimable with a large set of covariates, and because some teachers were not matchable with the full set of covariates using the independent PSAs, a smaller subset of covariates was identified that most impacted student final mathematics test scores, using stepwise regression. This smaller set of covariates was used as matching

variables in a second set of value-added analyses and independent propensity score analyses and within the multinomial models.

Analyses

The analysis encompassed four steps: (a) value-added analyses, (b) logit-model propensity score analyses done separately for each teacher, (c) generalized PSA analyses done separately by stratum, and (d) effect estimate comparisons. Each of these steps required a large number of smaller decisions and details.

Value-added models. Two value-added models were estimated in R in order to calculate value-added effects for the teachers. Both models were of the form:

$$\mathbf{Y}_{ij} = \mu + \boldsymbol{\tau}_j + \mathbf{X}_{ij}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{ij} \quad (2)$$

where i indexed students and j indexed teachers. $\mathbf{X}_{ij}$ was a matrix representing the same covariates used in the propensity score analyses, $\mathbf{Y}_{ij}$ represented the end of year kindergarten IRT mathematics ability estimate, and $\boldsymbol{\tau}_j$ represented the effect of teacher j. In the first model, $\boldsymbol{\tau}_j$ was treated as a fixed effect, and in the second model, $\boldsymbol{\tau}_j$ was random. In addition, each model was re-estimated using a selection of smaller subsets of covariates. For each model estimation, the teacher effects were ranked, organized into quintiles (as in Koedel & Betts, 2007; McCaffrey et al., 2009), and compared using Pearson and Spearman rank correlations and quintile comparisons.

The practice of ranking estimated effects into quintiles and assessing to what degree the effects move across quintiles using different model forms, which is often used to compare value-added estimates, had the advantage of providing practical information about how conclusions about teachers may change simply by altering the PSA methodology. If a large number of teachers in the lowest quintile using one method moved to a higher quintile when another

method was used, then making high stakes decisions about low-rated teacher would be more questionable.

In addition, the concordance correlation coefficient (r_c) was used for effect comparisons. The concordance correlation evaluated how closely the relationship between two measures fit a 45 degree line (intercept=0 and slope=1), suggesting they were identical (Lin, 1989).

Propensity score analyses. Four hundred thirty-five separate propensity score analyses were conducted in the statistical software package R (R Development Core Team, 2011), one for each teacher. Each analysis was done twice—once with the larger set of covariates and once with the smaller subset. All students in the teacher's class were considered the *treatment* group. This treatment group of students was matched to the best possible *control* group of students from others in the total sample of 11,451 students. This resulted in 435 treatment-group/control-group pairs. While each treatment group was unique (one teacher's class), the control groups may have overlapped in terms of students assigned. The goal was to find for each teacher's class of students the best possible control group in terms of the propensity score (probability of assignment to treatment).

Propensity scores were estimated using logistic regression (*glm* function in R), with the selected matching variables. Using these propensity scores, a control group of students was matched with each treatment group using optimal matching (*optmatch* package in R).

In optimal matching, network flow theory is used to select a control group with the minimum total difference between propensity scores of treated units and their matched control(s). However, there were a large number of options in optimal matching related to the number of controls to be assigned to each treated case. Rosenbaum (2002) and Guo and Fraser

(2010) both indicated that most propensity studies involve using several of these options and comparing results, and that practice was followed in this study.

Assignment of controls to treated subjects resulted in matched *sets,* with s_i treated units and t_i untreated units in each set *i* (notation used by Hodges & Lehmann, 1962). With the optimal matching methods used in this study, $s_i = 1$ for all *i*. In other words, a *set* consisted of one treated unit and one or more matched controls. If also $t_i = 1$, then we had matched pairs, sets with one treatment and one control. While a matched pairs design resulted in the minimum total difference in propensity scores, the goal of optimal matching procedures, there would have been a loss of information from other potential controls, if available. The several solutions to this conflict have resulted in the various matching options used in optimal matching. The goal was to waste as little information in the controls as possible while still making matches to the treated subjects that were as close as possible. One solution was to force a specified treated/control ratio, such as 1:2, 1:3, or higher. An advantage of this design was that it created balanced sets, which were simple to work with computationally. A disadvantage was that some treated units may have ended up with one or more poor matches, because the matches were forced. However, this risk is reduced when the ratio of potential controls to treated units is large, as in this study. Because the best approach has not yet been determined, and may depend on quirks in individual datasets, researchers typically have used several approaches and compared results.

The use of the large national dataset in this study meant that the number of potential controls for each treated unit (student in a particular teacher's class) was large. This increased the odds that multiple good matches could be made for any student. In this study, I created matched sets for the students in each teacher's classroom using the following rules:

1. 1:1 fixed ratio of treated unit (student in teacher's class) to control
2. 1:2 fixed ratio
3. 1:5 fixed ratio

A 1:20 fixed ratio was also used briefly for the examination of the relationship between matching quality and matching ratio. However, teacher effects were only estimated and compared for the 1:1, 1:2, and 1:5 ratios.

After the control sample was chosen for each teacher by each method, each teacher-control pair of samples was checked for balance on the matching variables using independent sample *t*-tests. If any of the variables had *t*-test statistics that were statistically significant ($\alpha = .05$) for a large number of teachers, a higher order term would have been added to the model. As Guo & Fraser (2010) suggested, there is no definitive procedure for creating the form of the model or selecting covariates. Sometimes balance cannot be improved simply because there is insufficient overlap between the distributions of covariates for the treated subject and the potential controls. The extremely large ratio of potential controls to the treated made this possibility seem unlikely in this study, but it became a problem. If adding square terms could not improve the balance of the treatment-control group pairs, then it was determined that the model was the best fitting possible, keeping in mind that 5% of covariates should have had statistically significant poor balance across any teacher-control group pair just due to chance. In addition, in this study match quality was evaluated by examining the difference between the propensity scores for all matched student-control pairs within a class.

For many teachers in the sample, when a large set of matching variables was used quality matches could not be made with the potential controls available. As a result, some comparisons were made with only the subset of teachers for whom matching was possible. In addition, a

second set of propensity score analyses was done for all teachers using a smaller set of covariates.

After matching, the *teacher effect*, was estimated using the identified control group (a) by comparison of means across each treatment and control groups, (b) by regressing the response on the dummy treatment variable (teacher assignment) for each class separately, using the mathematics pretest as a covariate, and (c) by adding a small selection of additional covariates to the regression models estimated in (b).

This process was repeated with each matching methodology for each of the 435 teachers, and resulting treatment effects were considered the *teacher effects*. Next, the 435 teacher effect estimates were ranked. The 435 effects were divided into five *quintiles,* or equal-sized groups, by rank. The teacher effects resulting from the various propensity score analysis methods were compared several ways: (a) a Pearson correlation was calculated, (b) a Spearman rank correlation was calculated for the two sets of ranked effects in order to get a single-value measure of how well the two sets of estimates were related as monotonic functions of each other, (c) the percent of teachers who move from one quintile to another using the two PSA methods was found, and (d) the concordance correlation coefficient was estimated.

Generalized propensity score analyses. Next, multinomial logit regressions were estimated for each stratum separately. For each stratum, a vector of propensity scores was estimated. Following Imbens (2000), the expected value of the response (end of kindergarten mathematics achievement) was estimated for each teacher, given the covariates (a dose-response function). After these expected values were estimated for all strata, teacher effects were calculated by subtracting each teacher's response from the overall mean of the expected responses for all teachers, across all strata. As previously described, only a random selection of

teachers within each stratum were included in each of the analyses, with the number of teachers per stratum selected dependent on how well the estimations converged.

Comparisons. Teacher effect estimate rankings found using each of methodologies were compared. As described above, the Pearson, concordance correlation, rank correlation, and quintile comparisons were used to make pairwise comparisons for the various VA and PSA sets of estimates.

In addition, the sets of PSA and VA effects were compared within pairs of classrooms with similar covariates. For each possible pair of classrooms, covariates were compared using the Mahalanobis distance. For each classroom, the most similar class was identified, and teacher effects estimated from each of the approaches were examined for the two classes in order to determine whether the ranking of the two teachers was consistent across the methods.

The hope was that this study would shed greater light on how data about teachers, students, and classroom compositions might best be used in order to estimate teacher inputs to student achievement for accountability purposes.

Chapter 4: Results

Analysis of the data proceeded as described in the methods section, beginning with selection of the sample and potential covariates, and proceeding through the various analyses and comparisons.

Sample Description

The final sample consisted of 435 teachers teaching a total of 4617 students. An additional 6834 students were included whose class sizes did not meet the selection criteria for the estimation of teacher effects but who were used as potential controls for the 435 classes of interest, resulting in a final pool of 11,451 students. For the 435 teachers, class sizes ranged from 5 to 25 students, with a mean of 10.61 students per class and a standard deviation of 4.79. For the 11,451 students, the mean mathematics test score gain over the kindergarten year was 10.26 units on the IRT-based scale, and the standard deviation was 6.96.

Covariate Selection

Originally, 187 potential child and family covariates were selected from the ECLS-K dataset. Since school and class variables were confounded with teacher assignment (only one teacher was selected from each school), they could not be used in the logistic regressions used for the propensity score analyses and so were not selected. The 187 variables were chosen based on their potential to relate to kindergarten mathematics test score gains. The variables included socio-economic measures and demographics, child assessments and development history, and family composition characteristics. A complete list of these initially selected variables is provided in Appendix B.

Next, variables that had very little variability or excess amounts of missing data within the 435 classrooms of interest were eliminated. Among the 105 categorical variables, 47 had at

least 80% missing data or at least 95% of the responses within one category and were eliminated. Similarly, among the 82 scale variables, 29 were found to contain very little data or near-zero variability, often because the variable tested a skill that was beyond the capacity of most kindergarteners, such as understanding place value, or applied to very few of the students, such as the number of weeks of prematurity at birth. The 111 remaining variables are listed in Appendix C.

These 111 remaining variables were examined for imbalance across the 435 classrooms, using one-way ANOVA for the scale variables and chi-square tests for the categorical variables. Variables with p-values greater than .30 (very little evidence of imbalance) were removed. This criterion was chosen conservatively as the goal was to keep variables which *may* have had imbalance. Additionally, variables that had high bivariate collinearity with others ($r \geq .80$) were removed at this stage. Two variables (birth weight pounds and birth weight ounces) were combined to create one variable. The resulting set contained 70 potential covariates. At this stage, missing data were imputed using maximum-likelihood from the *Amelia* package in R, accounting for nominal and ordinal variables.

While bivariate collinearity had been addressed earlier, multicollinearity had not. The variance inflation factor (Allison, 1999), or *VIF*, was found for each potential covariate after regressing it on the others. The largest *VIF* was found for C1R4MSCL, the beginning of year mathematics score. Theoretically, this variable could not be moved from the analyses as it formed the basis of the value-added analyses, and so those variables it was multi-collinear with, C1R4MPB1, C1R4MPB2, and C1R4MPB3, three other mathematics assessments, were removed instead. Two other variables, P1CARNOW and P1PRIMNW (two measures of current daycare arrangements) were found to be essentially the same and so the latter was removed. Sixty-six

covariates remained and were used as a starting point in the analyses. These covariates are listed and described in Appendix D.

Among these remaining variables, the highest *VIF* was now 1.99. Allison (1999) suggests that a *VIF* over 2.50 indicates potentially problematic multicollinearity. The set included 27 scale variables, 16 ordinal variables, and 23 nominal variables. As summarized in Appendix D, the variables represented student background experiences, parent variables, and measurements taken on students. The variables had not been, at this point, examined in terms of their actual impact on the outcome variable, the end of year mathematics test score.

Value-Added Models

For comparability with the propensity score models, both value-added models used all 11,451 students. In the fixed effects model, the students who did not belong to one of the 435 classes of interest were designated as belonging to class 436, and this group was used as the reference class. In the random effects model, the actual class assignment was modeled for all 11,451 students. In both cases, teacher effects were only extracted and compared for the 435 teachers of interest. Fixed and random effects models were estimated with successively smaller subsets of covariates (66, 17, 7, and 1). In all cases the beginning of year kindergarten mathematics exam was included. These variations served two purposes:

1. Comparability with the propensity score and generalized propensity score analyses, which in some cases were only estimable with fewer variables.
2. Examination of the consequences of eliminating covariates on the value-added teacher effect estimates.

R's *lm* function was used to estimate fixed teacher effects. Random effects were estimated using the *lmer* function within R's *lme4* package.

Sixty-six covariates. Using all 66 covariates, fixed teacher effects ranged from -11.63 to 9.07 with a mean of -0.47 and a standard deviation of 2.92. Random effects estimates ranged from -7.73 to 4.01 with a mean of -0.22 and a standard deviation of 1.59. For these two sets of estimates, both the Pearson product-moment correlation coefficient (r = .98) and the Spearman rank correlation coefficient (r_s = .99) were high. As shown in Figure 1, the two sets of teacher effect estimates were most similar to each other when they were closest to zero. This similarity at the center was expected due to the weighting or shrinkage of random effects estimates toward the mean, especially for classes which were small or had high within-class variability. The lower variability for the random than for the fixed effects resulted in a drop in the concordance correlation coefficient, however, in spite of the strong linear relationship between the two sets of estimates. The concordance correlation coefficient (r_c = .81) reflected the difference in both scale and rotation from the line of perfect concordance, as illustrated by Figure 1.

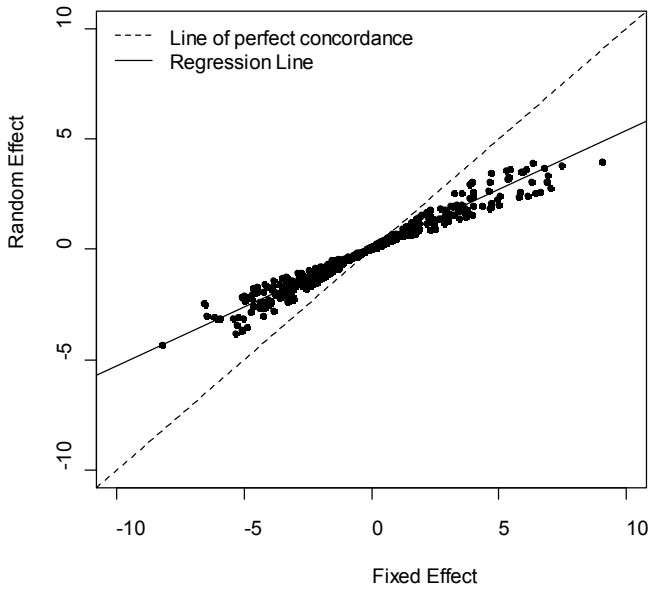


Figure 1. Concordance of fixed and random effects with 66 covariates. This figure illustrates the relationship between the fixed and random effect estimates.

When teachers were ranked and categorized into five quintiles twice, first using random effect estimates and then using fixed effect estimates, the teachers fell into the same quintile of rank 89% of the time. For example, of the 87 teachers who were ranked in the lowest 20% using fixed effects, 79 were also ranked in the lowest 20% using random effects, but eight teachers had moved up by one quintile. Those who differed in their quintile assignment across estimation methods never differed by more than one quintile (Table 1).

Table 1
Quintile Comparisons, Random and Fixed Effects with 66 Covariates

	Random Effect Counts				
Fixed Effect Counts	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Quintile 1	79	8	0	0	0
Quintile 2	8	72	7	0	0
Quintile 3	0	7	77	3	0
Quintile 4	0	0	3	78	6
Quintile 5	0	0	0	6	81

The random and fixed effects approaches appeared to rank the teachers in the sample similarly, although the two sets of estimates were not identical.

In these value-added models, only 11 of the included covariates had statistically significant ($\alpha = .05$) slope estimates in the presence of the other covariates (Appendix F). However, these covariate slope parameter estimates were nearly perfectly correlated ($r = .99$).

Seventeen covariates. Stepwise regression was employed for the fixed effects approach in order to identify a reduced model with good fit. This procedure became necessary later in order to identify the variables which best predicted the outcome for use in the propensity score and generalized (multinomial) propensity score models, which had estimation problems when all

66 covariates were included. Variables were removed one at a time using backward elimination, with the variables with the highest p-values removed first, provided there was no significant effect on the model R^2. Forty-nine covariates were removed, resulting in a minimal drop in model fit (R^2 = .690, 66-covariate model; R^2 = .689, 17-covariate model). This cutoff of 17 covariates also coincided with the largest set of covariates that made the multinomial models estimable. No covariate was removed with a p-value less than .10 in the step before exclusion. The final set of covariates is listed in Appendix E, Table E1. Estimates from this model correlated highly with estimates from the other value-added models, as shown in Table 2.

Seven covariates. Random and fixed effects models were estimated using the seven covariates which had the lowest p-values on the slope parameters in both the full (66 covariate) and reduced (17 covariate) value-added models. These seven covariates are listed in Appendix E, Table E2. The purpose of these models was to compare results with propensity score approaches that used a linear model with covariates at the effect-estimation stage. Because the smallest class included in the study had 5 members, and the smallest matching scheme in the propensity score analyses was one-to-one, only ten degrees of freedom were available in the smallest samples. Again, teacher effects from the seven-covariate value-added models were highly correlated with the other value-added models (Table 2).

One covariate. For both random and fixed effects approaches, baseline models with no covariates other than the beginning of kindergarten mathematics test (C1RSMSCL) were also estimated. Table 2 reflects the high correlation of effects from this model with the other value-added models. For both random and fixed teacher effect estimation approaches, the correlation between the two extremes—the full (all 66 covariate) and minimal (only one covariate) models—was very high (r = .95 random, r =.95 fixed). Concordance correlation coefficients

Table 2

Value-Added Effect Correlation Matrix using 1,7,17, and 66 Covariates

	Random Teacher Effect				Fixed Teacher Effect			
	Number of Covariates				Number of Covariates			
	1	7	17	66	1	7	17	66
Random Teacher Effect								
1 Covariate	-	.955	.951	.947	.982	.938	.934	.930
7 Covariates	.950	-	.998	.996	.935	.981	.979	.977
17 Covariates	.944	.997	-	.999	.931	.979	.981	.980
66 Covariates	.940	.995	.998	-	.927	.977	.979	.981
Fixed Teacher Effect								
1 Covariate	.992	.941	.935	.931	-	.953	.949	.945
7 Covariates	.945	.992	.989	.987	.949	-	.998	.996
17 Covariates	.940	.990	.992	.990	.944	.997	-	.998
66 Covariates	.936	.988	.991	.993	.940	.995	.998	-

Note. Pearson correlations above diagonal. Spearman correlations below diagonal.

suggested that teacher effects were essentially the same whether or not the remaining 65 covariates were included in the model (r_c = .93 random, r_c = .93 fixed). At the two extremes of covariate inclusion, the teachers' quintile rankings differed 32% of the time when random effects were estimated and 34% of the time when fixed effects were estimated. These rankings, however, rarely varied by more than one quintile (Table 3).

Table 3
Quintile Comparison Matrices, 66 and 1 Covariates

	Random Effect, 66 Covariates				
Random Effect, 1 Covariate	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Quintile 1	68	18	1	0	0
Quintile 2	18	47	20	2	0
Quintile 3	1	21	46	18	1
Quintile 4	0	1	19	57	10
Quintile 5	0	0	1	10	76

	Fixed Effect, 66 Covariates				
Fixed Effect, 1 Covariate	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Quintile 1	68	19	0	0	0
Quintile 2	17	46	23	1	0
Quintile 3	2	18	46	20	1
Quintile 4	0	4	17	53	13
Quintile 5	0	0	1	13	73

Figures 2 and 3 illustrate the high concordance between value-added teacher effects estimated with all 66 covariates and teacher effects estimated with only the mathematics pretest score used as a covariate for both fixed and random effects estimate approaches. In both cases, the teacher effect estimates with 66 and one covariates closely fit the line of perfect concordance.

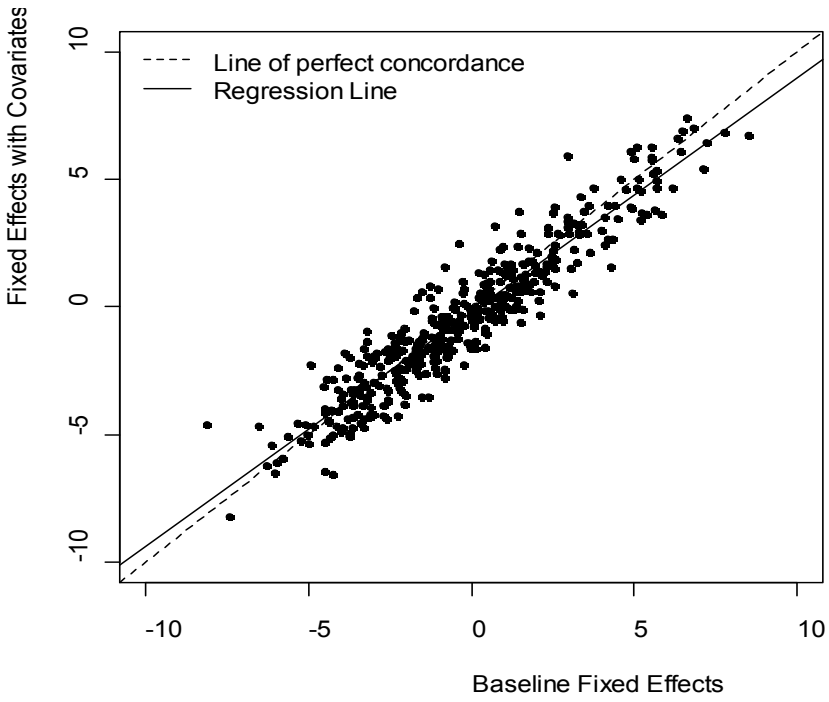


Figure 2. Concordance of fixed effects with 66 and 1 covariates.

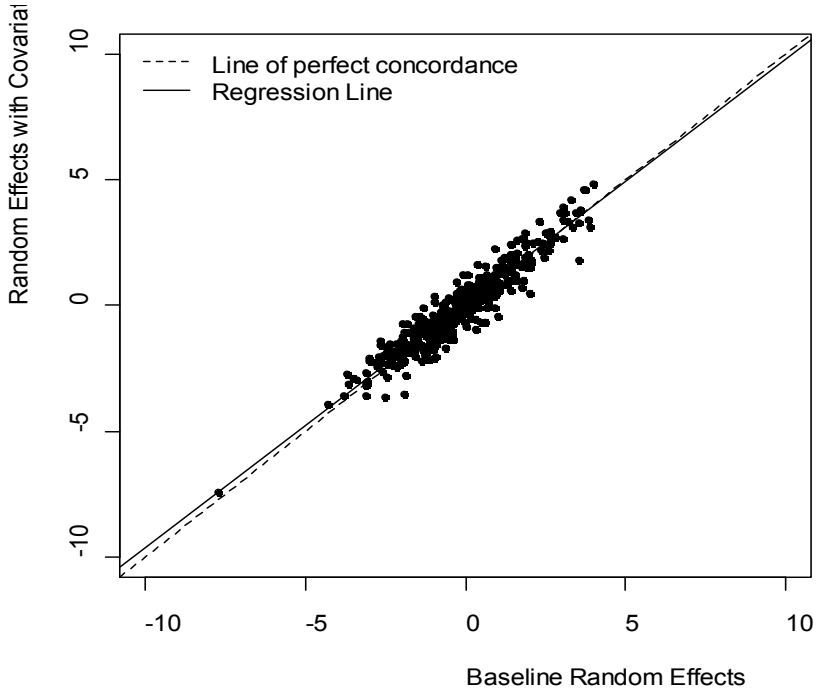


Figure 3. Concordance of random effects with 66 and 1 covariates.

Independent Propensity Score Models

Estimation of the propensity scores using the 435 independent logistic models resulted in one major complication. Even though the matching variables were chosen to exclude those with minimum variability overall, often the variables had minimal or no variability within specific classes. For 236 of the classes, there was perfect separability between the treated and potential controls when all 66 matching variables were used. Essentially, there was no region of common support. The set of matching variables was able to perfectly predict which students were in the teacher's class. The teacher's students received propensity scores of near-one and the others received propensity scores of near-zero. For each of these classes, the teacher's students were essentially unique given this set of matching variables. They could not be matched.

However, the sample sizes from these unmatchable classes (mean of 7.92 students) tended to be smaller than the sample sizes from the matchable classes (mean of 14.35 students). The sampling design unique to this study (random sampling of students at the within-school level) limited class sizes in a way that should not occur in a true teacher accountability data collection. As a result, a smaller proportion of teachers would likely be affected by separability, given covariate distributions similar to this study.

For this dataset, there were several possible solutions to the separability problem, all of which were attempted:

1. Proceed anyway, allowing matches to be created which were near-random and which did not improve teacher-control group balance on the covariates for those 236 unmatchable teachers. For these 236 teachers, effect estimates were analogous to unmatched comparisons as made in the value-added models.

2. Reduce the number of matching variables, especially those which may have had little effect on outcomes, so that the separation wasn't as strong. Many of the covariates, while they were related to classroom assignment and contributed to separability, were found in the value-added analyses not to contribute significantly to student outcomes. In other words, ignoring the sorting based on those characteristics did not bias teacher effect estimates significantly—it simply lost the ability to balance classrooms on unimportant characteristics. As a result, the set of 17 variables (see Appendix E, Table E1) found in the value-added models to best predict outcomes while still making the multinomial models estimable was used in one set of propensity score analyses. With these 17 variables, the extreme separation was not a problem for any remaining teachers.
3. Consider these 236 teachers' classes unique and, therefore, these teachers' true effects on student test scores relative to other teachers to be un-estimable. In order to examine the impact of the full set of 66 matching variables on teacher estimates in this study, the 236 unmatchable teacher effects were excluded for some of this study's comparisons.

Using the two sets of matching variables (66 and 17), 435 logistic regressions were estimated in R, using the *glm* function. Using each variable set, predicted values (propensity scores) were found for all 11,451 students for each teacher assignment. After these sets of propensity scores were estimated, R's *optmatch* package was used to match students in each of the 435 classes with one, two, or five other students in the remaining pool of students.

The approach in this project was to find the effect on each teacher's actual students of assignment to that teacher. For this reason, the balancing of the treatment-control group was

one-directional. Control groups were assigned that best balanced the characteristics of each teacher's students. No attempt was made to select for consideration the most appropriate students from within the teacher's class so that the class better reflected the balance of characteristics in larger student population. All of the teacher's students with available data were used. In other words, the interest was in the average treatment effect on the treated, the effect on each teacher's actual students of assignment to that teacher.

Match quality. After matching, each class-control group was checked for match quality and improved covariate balance. A *t*-test comparing either means or proportions was conducted for every variable, for every class-control pair, for every matching scenario. Class-control pairs for which the difference in means or proportions was statistically significant at $\alpha = .05$ for a given variable were tallied for each scenario. For comparison, identical *t*-tests were conducting comparing the mean or proportion for each variable for each teacher before matching, using all potential controls as the comparison group. Table G1 in Appendix G lists the number of teachers for whom the imbalance on a variable was statistically significant both before and after matching. On all variables, the number of class-control pairs with statistically significant imbalance dropped dramatically after matching, with the number remaining less than the 5% expected due to chance at $\alpha = .05$. Statistically significant imbalance was decreased in most classrooms in spite of the separability (unmatchability) problem for over half the teachers, though to some degree this was expected due to a decrease in statistical power.

For comparison, balance after matching was also examined for a 1:20 match ratio with the expected result that more teacher-control pairs remained statistically out of balance under that scenario. Table G2 (Appendix G) repeats the procedure with the 17 matching variable model. Again, little imbalance remained after matching.

With the 17 matching variable model, the number of class-control pairs with statistically significant differences appeared to decrease on most variables as the matching ratio increased from 1:1 to 1:2 to 1:5, in spite of increased power to find differences with the larger ratios. With the 66 matching variable model, results were less consistent, and the 1:2 matching ratio appeared to indicate the fewest classrooms with remaining imbalance across most variables.

In addition, the matched distances for all student-control pairs were examined. Using the results from the 66-variable model, median matched distances for each of the 199 matchable classes were found. Because the matched distances for the classes suffering from the separation problem were always almost 1.0, they were removed for this analysis. The goal was to find out how close the matches were for those teachers who *could* be matched. For each class, the median class matched distance was found. Table 4 shows the mean and median across all classes of these median matched distances as the matching ratio was allowed to increase. Again for comparison, matching distances were also found for a 1:20 matching scheme. Matches decreased in quality as more matches were found for each student.

Table 4
Median Matched Distance, 66 Covariates, 199 Matchable Teachers

Matching Ratio	Median[a]	Mean[b]	SD[c]
1:1 Match	0.001	0.007	0.018
1:2 Match	0.003	0.013	0.028
1:5 Match	0.010	0.031	0.047
1:20 Match	0.087	0.126	0.118

[a]Median of class medians.
[b]Mean of class medians.
[c]SD of class medians.

The same procedure was followed for the 17 matching variable approach, this time using all 435 teachers (Table 5). Comparison of Tables 4 and 5 demonstrates that the matching quality was noticeably improved when only 17 variables were included in the model. This reinforces the idea that the excluded variables posed a matching problem for the teachers. Even when only the 199 matchable teachers were included in the analysis, the matching quality for those teachers was poor using 66 matching variables compared with the 17 matching variable model. Essentially, the extra matching variables appeared to be requiring balance on class-control pairs that could not be achieved well with the sample of 11,451 potential controls available.

Table 5
Median Matched Distance, 17 Covariates, All Teachers

Matching Ratio	Median[a]	Mean[b]	SD[c]
1:1 Match	0.001	0.001	0.003
1:2 Match	0.001	0.002	0.007
1:5 Match	0.001	0.004	0.016
1:20 Match	0.003	0.018	0.066

[a]Median of class medians.
[b]Mean of class medians.
[c]SD of class medians.

Teacher effect estimation. Teacher effects were estimated three different ways:

1. Difference in means: The difference in estimated mean end of year mathematics test score between the teacher's students and the students matched to the group was estimated. This effect estimation approach did not take into account any of the covariates at the effect estimation stage, but instead assumed that accounting for covariates at the propensity score estimation and matching stage was sufficient.

2. Slope with pretest: The estimated slope on the teacher dummy variable was found by estimating a linear model with end of year mathematics test as the dependent variable and the mathematics pretest as a covariate. As a result, the fall mathematics test (pretest) was used twice—once at the propensity score estimation stage and again when teacher effects were estimated. This approach was analogous to estimating successive simple value-added models for each teacher separately with a reduced sample including only each teacher's students and assigned controls.
3. Slope with multiple covariates: The estimated slope on the teacher dummy variable was found by estimating a linear model with end of year mathematics test as the dependent variable and a small set of covariates. Seven variables that had p-values less than .01 in the VA fixed regressions were included (see Appendix E, Table E2). This set of covariates was intentionally kept small as the smallest classrooms included only five students and the smallest matching ratio was 1:1, meaning treatment-control group combined samples had a lower limit of $n = 10$.

For each of these three effect estimation methods, each of the three approaches to the separability (unmatchable teacher) problem addressed earlier was used: (a) All teacher effects were estimated using all 66 variables, allowing unmatchable teachers to be matched to controls almost randomly, (b) all teacher effects were estimated using a reduced set of 17 covariates, and (c) only the matchable teachers using 66 variables were included. In addition, each of the three matching schemes (1:1, 1:2, and 1:5) was used.

Sixty-six variables, all teachers. Within the 66-variable all-teacher design, the three effect estimation methods and the three matching schemes produced 435 teacher effect estimates with Pearson product-moment correlations ranging from $r = .61$ to $r = .98$, as shown in Table 6.

Table 6
Correlation Matrix for Matching and Effect Estimation Schemes when 66 Matching Variables and All Teachers were Used

	Means			1 Covariate			7 Covariates			Value-Added	
	1:1	1:2	1:5	1:1	1:2	1:5	1:1	1:2	1:5	Fixed	Random
Means											
1:1 Match	-	.81	.70	.82	.74	.72	.68	.71	.71	.67	.65
1:2 Match	.81	-	.85	.76	.88	.83	.64	.83	.81	.78	.76
1:5 Match	.71	.84	-	.72	.81	.89	.61	.76	.86	.84	.83
1 Covariate											
1:1 Match	.81	.75	.72	-	.88	.83	.83	.84	.84	.78	.76
1:2 Match	.73	.87	.80	.86	-	.92	.76	.95	.92	.86	.84
1:5 Match	.71	.81	.88	.81	.90	-	.70	.87	.98	.94	.92
7 Covariates											
1:1 Match	.72	.66	.64	.88	.76	.71	-	.80	.72	.67	.64
1:2 Match	.71	.82	.76	.83	.95	.85	.81	-	.89	.81	.79
1:5 Match	.70	.79	.85	.82	.90	.98	.74	.87	-	.93	.91
Value-Added											
Fixed	.67	.75	.82	.76	.84	.94	.68	.80	.92	-	.98
Random	.66	.75	.81	.76	.84	.93	.68	.79	.91	.99	-

Note. Pearson correlations above diagonal. Spearman correlations below diagonal.

Both the matching ratio and the effect estimation approach impacted the correlations between the estimates of teacher effectiveness.

In all cases, as the matching ratio increased, the correlation of the propensity score-based effects with the value-added effects increased. Effects estimated using identical matching ratios were always more highly correlated than those using different ratios, regardless of the effect estimation method used (Table 6). One-covariate estimation models were more correlated with the value-added effects, however, than were the seven-covariate estimates, though both sets of correlations were higher than those resulting from the mean-difference estimates.

All else being held constant, varying the level of covariate inclusion (mean differences, 1 covariate, 7 covariates) at the effect estimation stage produced Pearson correlations that were large ($r > .68$). These correlations were higher when a larger number of imprecise matches were assigned to each student rather than a fewer number of more precise ones (Table 6). Although these correlations were large, the value-added estimate correlations had been much stronger ($r > .945$) regardless of the level of covariate inclusion (Table 2). The propensity score-based estimates were more sensitive to the level of covariate inclusion at the effect estimation stage than were the value-added estimates.

For these 435 teachers, effect correlation across propensity score approaches was weaker than across value-added approaches. While a correlation of $r = .61$ may be considered high within other contexts, when high stakes decisions are being made about teachers, that level of correlation is likely insufficient. A correlation of $r = .61$ suggests that only about 37% (r^2) of the variability in one set of teacher effects can be predicted, or explained, by the other set of teacher effects.

Table 7 shows the concordance correlation coefficient (r_c) showing the relationships between sets of estimates. These values were generally very close to the Pearson correlations. Concordance correlations with the random value-added estimates, however, were lower because r_c takes into account departures from the slope of the line $y = x$ in addition to the tightness of fit to the line. The reduced variability at the extremes when estimating random effects alters the slope of this line.

Propensity score-based effect estimates could not be estimated with more than a few covariates due to degrees of freedom limitations, as described earlier. All comparisons with propensity score estimates, thus far, had used estimates from the full 66-variable value-added model. These value-added 66-covariate estimates were very highly correlated with value-added estimates using fewer covariates ($r > .945$). However, in order to more carefully examine the relationship between propensity score matching ratios and value-added estimates, estimates from the fixed effects value-added model with only seven covariates were correlated with propensity score estimates using the same seven covariates for effect estimation and 1:1, 1:2, 1:5, and 1:20 matching schemes. As illustrated in Table 8, the correlation between the value-added and propensity score-based estimates appeared to converge to 1.0 as the matching ratio increased, all else being held constant. Theoretically, a one-to-all propensity score-based matching scheme should duplicate a fixed effects value-added model, if identical covariates are included in the effect estimation (linear model) stage as in the value-added model. In each case all students not assigned to the teacher's class would be used as controls.

While correlations indicate the strength of relationship between two sets of estimates, they do not indicate what practical effect any differences in estimates would have on high stakes

Table 7

Concordance Correlation Coefficient Matrix for 66 Matching Variables and 435 Teachers

	Means		1 Covariate			7 Covariates			Value-Added	
	1:2	1:5	1:1	1:2	1:5	1:1	1:2	1:5	Fixed	Random
Means										
1:1 Match	.80	.68	.79	.71	.66	.68	.69	.66	.60	.39
1:2 Match		.84	.75	.87	.80	.63	.83	.79	.75	.53
1:5 Match			.72	.81	.88	.58	.76	.85	.82	.62
1 Covariate										
1:1 Match				.87	.81	.80	.84	.82	.75	.54
1:2 Match					.91	.72	.95	.91	.84	.63
1:5 Match						.63	.86	.98	.94	.75
7 Covariates										
1:1 Match							.72	.63	.59	.39
1:2 Match								.86	.79	.75
1:5 Match									.93	.73
Value-Added										
Fixed										.82

Table 8
Correlations of VA Fixed Effects (7 Covariates) with Propensity Score Effects (7 Covariates at Effect Estimation Stage)

Matching Ratio	*r*
1:1 Match	.67
1:2 Match	.81
1:5 Match	.93
1:20 Match	.98

decisions about teachers. The percent of teachers who move to a different quintile of rank due to matching and estimation differences is reflected in Tables 9, 10 and 11. Within the propensity score-based approaches, between 41% and 68% of teachers remained within the same quintile when one or more factors (matching ratio, effect estimation approach) was changed (Table 9). Between 81% and 98% of teachers did not shift by more than one quintile (Table 10). Up to 11% of the teachers moved from the bottom two quintiles to the top two quintiles, or vice versa, depending on the propensity score matching ratio and estimation scheme chosen (Table 11).

Figure 4 provides a visual example of the relationship between propensity score-based and fixed value-added estimates when all 66 covariates and all classes are included in both models. The two sets of estimates have a moderately strong linear relationship with some rotation from the line of perfect concordance, reflecting greater variance in the propensity score-based estimates. Of greatest concern to teacher accountability systems are those teachers ranked above the mean by one statistical procedure and below the mean by another, those falling in the second and fourth quadrants on the graph. Figure 4 indicates a small set of teachers falling within those quadrants.

Table 9
Matrix of Percent of 435 Teachers Falling within the Same Quintile across Effect Estimation Approaches and Matching Ratios Using 66 Matching Variables

	Means		1 Covariate			7 Covariates			Value-Added	
	1:2	1:5	1:1	1:2	1:5	1:1	1:2	1:5	Fixed	Random
Means										
1:1 Match	.52	.42	.53	.44	.43	.49	.44	.44	.40	.39
1:2 Match		.55	.50	.59	.53	.46	.56	.51	.44	.44
1:5 Match			.46	.52	.61	.41	.50	.59	.53	.51
1 Covariate										
1:1 Match				.59	.52	.68	.59	.53	.47	.48
1:2 Match					.62	.52	.76	.63	.55	.54
1:5 Match						.46	.57	.84	.68	.68
7 Covariates										
1:1 Match							.52	.46	.42	.42
1:2 Match								.60	.53	.54
1:5 Match									.66	.66
Value-Added										
Fixed										.89

Table 10
Matrix of Percent of 435 Teachers Falling within the Same or Adjacent Quintile across Effect Estimation Approaches and Matching Ratios Using 66 Matching Variables

	Means		1 Covariate			7 Covariates			Value-Added	
	1:2	1:5	1:1	1:2	1:5	1:1	1:2	1:5	Fixed	Random
Means										
1:1 Match	.92	.84	.90	.86	.81	.84	.84	.81	.80	.81
1:2 Match		.93	.85	.93	.89	.81	.90	.87	.86	.86
1:5 Match			.83	.89	.95	.81	.88	.93	.91	.91
1 Covariate										
1:1 Match				.92	.88	.94	.90	.88	.84	.85
1:2 Match					.96	.85	.98	.97	.92	.92
1:5 Match						.82	.93	1.00	.99	.99
7 Covariates										
1:1 Match							.88	.84	.81	.80
1:2 Match								.95	.90	.89
1:5 Match									.98	.98
Value-Added										
Fixed										1.00

Table 11
Matrix of Percent of 435 Teachers Falling Moving from the Bottom 2 Quintiles to the Top 2 Quintiles or Vice Versa across Effect Estimation Approaches and Matching Ratios Using 66 Matching Variables

	Means		1 Covariate			7 Covariates			Value-Added	
	1:2	1:5	1:1	1:2	1:5	1:1	1:2	1:5	Fixed	Random
Means										
1:1 Match		.08	.06	.08	.09	.09	.09	.09	.09	.09
1:2 Match		.04	.08	.03	.07	.11	.06	.08	.07	.07
1:5 Match			.10	.06	.02	.10	.08	.03	.05	.04
1 Covariate										
1:1 Match				.04	.06	.03	.05	.06	.07	.07
1:2 Match					.02	.08	.01	.03	.04	.04
1:5 Match						.09	.03	.00	.01	.01
7 Covariates										
1:1 Match							.06	.08	.09	.09
1:2 Match								.03	.05	.05
1:5 Match									.01	.01
Value-Added										
Fixed										.00

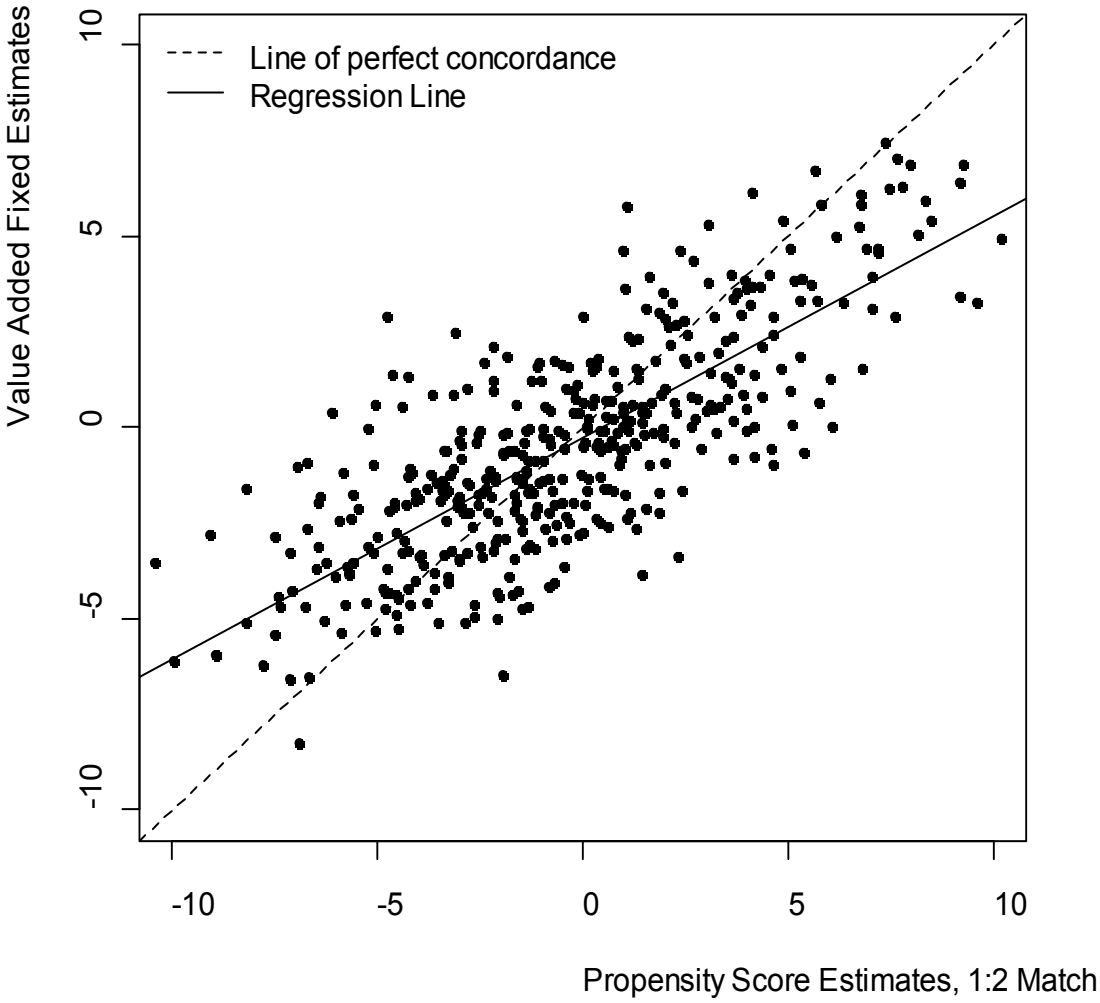


Figure 4. Diversion from line of perfect concordance of fixed VA estimates and propensity score estimates. All teachers, 66 covariates, a 1:2 matching scheme, and difference in means effect estimation, were used.

Seventeen variables, all teachers. Comparisons so far have involved propensity score estimates from a logistic model using all 66 covariates and all 435 teachers, in spite of the separability, or unmatchability, of many of the teachers' student sets. For these students, the matches that were made were essentially random. As a second approach, a reduced logistic model (17 matching variables) was estimated, eliminating the separability problem while retaining the variables that most impact student mathematics test scores. The resulting 435 teacher effects were correlated with similar estimates from the full model (66 matching variables) as shown in Table 12. The matching ratio had a large impact on the correlation of the effect estimates resulting from the full and reduced logistic models, regardless of the effect estimation approach chosen. The higher matching ratio may have replicated the poor matching that occurred for many teachers when the full logistic model was employed. The population from which the potential controls were selected was not large enough to provide precise matches when either a high matching ratio was chosen or too many matching variables were used.

Correlations within the 17-matching variable models and with the value-added models were slightly lower than those resulting from the problematic 66-covariate models (Table 13). The decreased linear relationship between the 17-matching variable model and the value-added model provides further evidence that value-added models are analogous to poorly matched propensity score models. Better matching of teachers' students to controls with the 17-variable model resulted in estimates that were less well matched to the value-added estimates (Table 6). However, it was not just the correlations with the value-added models that were lower for the 17 covariate case, but the correlations of effect estimates within the 17-covariate schemes as well (Table 13). The 17 variable models had more precise matches than the 66 variable models, which could not match some teachers' students.

Table 12
Correlations of Propensity Score-Based Effects with 66 and 17 Matching Variables, All Teachers

Matching Ratio	*r*
Means	
1:1 Match	.34
1:2 Match	.60
1:5 Match	.70
One Covariate	
1:1 Match	.57
1:2 Match	.75
1:5 Match	.87
Seven Covariates	
1:1 Match	.24
1:2 Match	.64
1:5 Match	.84

Table 13
Correlation Matrix for Matching and Effect Estimation Schemes when 17 Matching Variables and All Teachers were Used

	Means			1 Covariate			7 Covariates			Value-Added	
	1:1	1:2	1:5	1:1	1:2	1:5	1:1	1:2	1:5	Fixed	Random
Means											
1:1 Match	-	.78	.64	.76	.66	.60	.49	.60	.60	.58	.57
1:2 Match	.77	-	.85	.72	.84	.78	.47	.80	.76	.72	.71
1:5 Match	.64	.85	-	.69	.83	.92	.40	.75	.90	.85	.83
1 Covariate											
1:1 Match	.75	.74	.69	-	.85	.76	.62	.74	.74	.73	.72
1:2 Match	.63	.84	.83	.85	-	.91	.50	.90	.89	.85	.84
1:5 Match	.58	.77	.91	.74	.90	-	.42	.83	.99	.92	.91
7 Covariates											
1:1 Match	.63	.62	.58	.85	.72	.62	-	.52	.41	.41	.40
1:2 Match	.61	.80	.79	.81	.96	.86	.75	-	.83	.77	.76
1:5 Match	.58	.76	.89	.73	.89	.99	.61	.87	-	.92	.90
Value-Added											
Fixed	.55	.70	.83	.70	.84	.92	.59	.80	.91	-	.98
Random	.55	.71	.84	.70	.83	.91	.59	.80	.91	.99	-

Note. Pearson correlations above diagonal. Spearman correlations below diagonal.

Sixty-six variables, matchable teachers only. The third approach to dealing with the separability problem was to use only the 199 teachers for whom good matches could be obtained. When the effects from only the 199 matchable teachers from the 66-variable model were correlated, results were slightly higher than when the effects from all teachers were included either from the 66-covariate or 17-covariate model (Table 14). Removing the problem teachers made the results slightly more stable and consistent across matching ratios and effect estimation approaches.

Table 14
Pearson Correlation Matrix for Matching and Effect Estimation Schemes when 66 Matching Variables and 199 Teachers were Used

	Means		1 Covariate			7 Covariates			Value-Added	
	1:2	1:5	1:1	1:2	1:5	1:1	1:2	1:5	Fixed	Random
Means										
1:1 Match	.81	.70	.79	.72	.67	.76	.71	.67	.65	.65
1:2 Match		.88	.77	.87	.82	.72	.85	.81	.78	.79
1:5 Match			.74	.84	.91	.70	.82	.90	.87	.87
1 Covariate										
1:1 Match				.89	.83	.94	.88	.83	.79	.79
1:2 Match					.92	.84	.98	.92	.87	.87
1:5 Match						.78	.91	.99	.95	.95
7 Covariates										
1:1 Match							.86	.78	.75	.75
1:2 Match								.92	.86	.86
1:5 Match									.95	.95
Value-Added										
Fixed										.99

The relationship between the 66 and 17 covariate models was stronger when only the 199 matchable teachers were included (Table 15) than when all teachers were included (Table 12).

Table 15
Correlation of PSA Effects with 66 or 17 Matching Variables, 199 Teachers

Matching Ratio	*r*
Means	
1:1 Match	.323
1:2 Match	.558
1:5 Match	.726
One Covariate	
1:1 Match	.644
1:2 Match	.783
1:5 Match	.893
Seven Covariates	
1:1 Match	.622
1:2 Match	.769
1:5 Match	.896

Generalized Propensity Score Analyses

With this dataset, the multinomial models were only estimable if the combination of number of classes used in the model and the number of matching variables included was below some threshold, which varied by the classes and variables used. Dividing the 435 classes into 12 strata created strata sizes varying from 20 to 48 classes, not including the reference class used in the value-added and independent logistic models—the remaining students from the pool of 11,451 who were not members of any of the 435 classes of interest. In order to make the counterfactuals comparable across approaches, it was necessary to include all students in each

multinomial estimation. Including all students, moreover, increased the sample size and reduced some of the estimation problems.

Preliminary model estimations using a random selection of ten classes, plus all other students merged into a reference class, allowed the models to be estimated with about ten matching variables. Reducing the number of classes per analysis increased the number of matching variables that could be included without making the model unestimable. However, no matter which sets of classes and matching variables were selected, the estimated propensity scores using the multinomial approach correlated at $r > .999$ with the propensity scores found using the independent logistic models for the same classes and sets of matching variables.

Using the set of 17 variables which were found to have significant predictive impact on the outcome variable at $\alpha = .10$ in the value-added analyses allowed all multinomial models to be estimated when the 12 strata were divided into sub-strata of five or six classes, a total of 83 sub-strata. As found in the experimental analyses, correlations between independently estimated and generalized propensity score estimates exceeded $r = .999$. Whether estimated jointly or independently, the propensity scores were essentially identical, suggesting the easier to estimate independent models were more useful. While impossible to test at the extreme of large strata, it appears that propensity scores are nearly identical regardless of stratum size.

Imbens' (2000) dose-response function approach to effect estimation was employed by regressing outcomes on both generalized propensity score vectors and treatment assignment variables and then averaging across the covariate (propensity score) distributions to find the expected values of the response at each treatment level (class). Teacher effects were estimated by finding differences from the overall mean of expected values. These effects ranged from -12.06 to 6.73. The mean was 0.00 by design and the standard deviation was 2.29. These

teacher effects were essentially uncorrelated with all other sets of teacher effects estimated in this study ($r < .10$ for all effect pair).

It is important to note that the first step in Imbens' (2000) effect estimation procedure, the regression estimation, is analogous to fitting a fixed effects value-added model with 17 covariates. However, using Imbens' approach 435 non-linear combinations of the 17 covariates (the propensity scores) replaced the 17 covariates used in the value-added model. In other words, as described by Lechner (1999), the use of a multinomial approach increased, rather than decreased, the dimensionality of the problem. This complication suggests that this multinomial set of effect estimates is inferior to, rather than simply different than, the other sets of teacher effect estimates.

One compromise that did not increase the dimensionality of the problem was to find predicted values using the value-added regression model, then use Imbens' (2000) next step of weighting the predicted values using the propensity scores for each class. The theoretical implications of using this procedure were beyond the scope of this project, but the method clearly would not solve the problem of accounting for any non-linearities in the relationships between the covariates and the outcome. However, this compromise effect estimation approach was conducted. Resulting teacher effects were highly correlated with those found by Imbens' approach ($r = .90$), and therefore uncorrelated with all other sets ($r < .11$ for all effect pair).

Another option was to estimate and rank teacher effects within, rather than across, strata or sub-strata. While this approach would have aligned most closely with Imbens' (2000) theoretical framework and would have avoided the increased-dimensionality problem, it would have made teacher effects across strata or sub-strata less comparable.

Class-Level Matching

For each class, a mean for each of the 66 covariates was estimated. Each class was then matched to the most similar class relative to those covariates, using the Mahalanobis distance. After matching, the two classes were ranked according to the mean end of year mathematics test score. Within-pair rankings for these class sets were also found for teacher effect estimates from the value-added and propensity score-based approaches. The teacher's ranking using class-level matching was essentially uncorrelated with the rankings from the other approaches (Table 16).

Table 16
Class Level Matched Rankings and Other Model Rankings

Effect Estimation Method	Correlation
Difference in Means	
1:1 Match	.042
1:2 Match	.042
1:5 Match	.001
One Covariate	
1:1 Match	.009
1:2 Match	.010
1:5 Match	.013
7 Covariates	
1:1 Match	.050
1:2 Match	.017
1:5 Match	.001
Value-Added	
Fixed	.017
Random	.010

Chapter 5: Conclusions

The goal of this study was to compare various approaches to estimating teacher contributions to student test score gains. Value-added modeling, the approach increasingly used in accountability systems in the United States, has been criticized for a variety of reasons (Baker et al., 2010; Newton et al., 2010; Rothstein, 2009). Of particular concern in this study was potential bias due to the teacher effect estimates resulting from violations of the regression linearity assumption (Cochran & Rubin, 1973; Stuart, 2007), and inconsistent covariate distributions across groups that might increase the impact of bias due to some types of nonlinearities (Rubin, 2001; Stuart, 2007). Propensity score-based matching techniques were examined as alternatives potentially resolving these regression-related problems.

Reflections on Findings

The results of this study have implications regarding the types of treatment effects that can most easily be estimated for teachers. These effect types vary across methodologies—value-added modeling, propensity score analysis, and generalized propensity score analysis.

Treatment effects. In order to determine whether a specific effect estimate is biased, the first step is to determine what kind of effect is desired, or was obtainable given data limitations. Most often, the average treatment effect (ATE) is the goal. The ATE is the expected change to the outcome variable, if treatment rather than non-treatment had occurred, for the entire population of interest. In the setting of teacher accountability, the ATE refers to the effect a particular teacher would have had on the outcomes of *any* student if that student had been assigned to that teacher instead of to an average teacher. A second type of effect, the average treatment effect for the treated (ATT), is more limited. It refers to the effect the treatment had on those who actually received treatment. In the context of this study, it refers to the effect a

particular teacher assignment had on the *actual* students assigned to that teacher. In some contexts, and some may argue teacher accountability is one of those settings, the ATT is preferred to the ATE. However, more often the ATE is the goal.

All three methods used in this study—value-added regression methods, propensity score matching methods, and multinomial-based dose-response approaches—have potential to measure the ATE if the correct assumptions are met. With regression methods, the linearity assumption is critical. Unless expressly modeled otherwise, it is assumed that the true effect is constant across all levels of the covariates in the regression model. If this is not the case, then the ATE estimate is biased to some degree or other at various levels of the covariates. Unless the true ATE is constant, the estimated ATE is misleading. It is simply the best-fitting single value given the data. When the impact of a teacher on outcomes differs across levels of some covariate, however, a teacher assignment-covariate interaction term may be added to the model. This term adds complexity to interpretation. The teacher would have a different effect on outcomes for different covariate levels. Basing an accountability system on such a model would be problematic.

If it is possible to create treatment and control groups that are balanced on the covariates using propensity score analysis (PSA), the ATE can be estimated. When the covariate distributions of the two groups overlap, individuals from each group can be matched in such a way that the two distributions balance better than they did originally. Because the full covariate distributions from both groups were represented, the estimated effect is the ATE. Essentially, the teacher's effect is averaged across the covariate distribution as it exists in the entire population.

PSA estimates, however, are not unbiased estimates of the ATE if the covariate distributions of the treated and controls do not overlap—if a region of common support does not exist. To be precise, covariate combinations (subjects) encountered among either treatments or controls must have positive probability both of existing in the treatment group and of existing in the control group (Caliendo & Kopeinig, 2005). If not, an unbiased ATE cannot be estimated. Practically speaking, it means that all propensity scores for both treated and control individuals must be greater than zero and less than one. Covariate overlap must exist and balancing must be possible in both directions. There must not be individuals in either group who were so unique that they could not have existed in the other group.

In the current study, this condition was violated for many teachers' classes, no matter which sets of matching covariates were used. An unbiased ATE could not be estimated for most teachers using propensity score approaches, with the given covariates and sample sizes. The ATE might have been estimated for these teachers using value-added methods only if the covariates that were not balanced across classes had no effect on the (presumably fixed) teacher effect estimates.

Using PSA methods, however, an unbiased ATT can often be estimated even when the ATE cannot. If all propensity scores in the treated group are less than 1.0, then the ATT can be estimated, regardless of whether any propensity scores in the control group were equal to 0.0 (Caliendo & Kopeinig, 2005). Essentially this means that all members of the treatment group must be matchable to members of the control group, but the reverse does not need to hold true. When this one-way matching is possible, then the effect estimated is the effect of treatment on those actually assigned to the treatment group, the ATT.

In the current study, this more limited one-way balancing was possible for all teachers when 17 matching variables were used, but it was impossible for over half the teachers when 66 matching variables were used. The 17-variable model, therefore, estimated the ATT, if and only if the true treatment effect was independent of any variables not included in the model. Likewise, the 66-variable model estimated an unbiased ATT for the subset of 199 teachers, provided no other important covariates were missing. Essentially, in both cases the ATT was unbiased if the ignorable treatment assignment assumption was not violated. It was not possible to test the effect of covariates that were not measured or included in the dataset that may have affected outcomes. It was possible, however, to compare the 17 and 66 covariate model estimates for the 199 teachers for whom the ATT was estimable. As shown in Table 15, the correlations between those two sets of effects ranged from $r = .323$ to $r = .896$. Apparently, the exclusion of the extra 49 variables when the 17-variable model was estimated did affect results, suggesting a potential violation of the ignorable treatment assignment assumption.

While it is impossible to know which set of effect estimates more closely represented *truth*, it is clear that using all 66 covariates and all teachers resulted in a major assumption violation, the requirement for a one-directional region of common support, for more than half the teachers, as needed to estimate the ATT. This requirement did not appear to be violated for either the 17-covariate, 435 teacher estimations or the 66-covariate, 199 teacher approach. In choosing between these two sets of estimates, the major remaining considerations would be to evaluate which approach (a) was less likely to violate the ignorable treatment assignment assumption and (b) resulted in better matching quality. Since there is no penalty or assumption violation for including extra variables in the model, the tendency would be to trust those from the 66-covariate model, leaving the effects of the remaining 236 teachers un-estimated, provided

match quality was sufficient. However, the match quality was better using the 17-covariate model. The solution to this dilemma is straightforward—use a larger sample and as many covariates as possible.

In summary, a teacher accountability system wishing to use propensity score approaches in order to achieve unbiased treatment affects needs to keep in mind the following:

1. Violations of the ignorable treatment assignment assumption will be likely, and teacher effects will be biased if important covariates are omitted.
2. The student population included needs to be large for unbiased ATTs to be estimated for all teachers. The more covariates that are necessary, the larger this student population needs to be.
3. It is unlikely that the ATE will be estimable for most teachers, no matter what covariates or student population pool are available. Small classroom sample sizes prevent classroom covariate distributions from sufficiently covering the overall population covariate distribution.

Value-added models. The results of this study indicated that propensity score-based effect estimates were correlated with value-added estimates, but not as highly as value-added estimates from various models were correlated with each other. These reduced correlations suggested, but did not prove, that nonlinearities did exist in the relationships between the covariates and the response variable, potentially biasing the regression-based estimates. While a PSA approach averages across either the population covariate distribution (the ATE) or the treatment group covariate distribution (the ATT), the regression approach simply assumes that the teacher effect is fixed across levels of that variable (unless specifically modeled otherwise), and so does not average across either distribution.

In this sample, value-added estimates were fairly stable across both effect estimation approach (random or fixed) and covariate inclusion choice (1, 7, 17, or 66). This stability provides evidence of the reliability of value-added estimates for these teachers. All sets of value-added estimates apparently measured the same thing, and that measurement was nearly independent of the student characteristics used as covariates. However, this high correlation provided no evidence of the validity of value-added estimates. The fact that the value-added approaches all appeared to measure the same thing does not imply that they measure the *right* thing.

Previous research has frequently found that linear models including covariates other than pretest scores provide limited additional information beyond what was contained in the pretests themselves (Harris & McCaffrey, 2010; Levine & Painter, 2010; Lockwood et al., 2007; Schochet & Chiang, 2010). The implication has been that adding additional covariates to the model is unnecessary when enough years of prior test scores are included. However, one piece of evidence suggesting that pretest scores do not simply duplicate the information found in the other covariates in the dataset was that the mathematics pretest score was not highly collinear or multi-collinear with the other covariates ($VIF = 1.99$), unless other mathematics tests were included in the model ($VIF = 3.49$).

It is possible that additional covariates do provide value to teacher effect estimation, but not when those effects are modeled linearly. In other words, it may that covariates do not significantly alter linear model-based estimates because something is wrong with the linear model, rather than with the covariates. While linear models are generally easy to estimate, it is common to misinterpret the resulting estimates when key assumptions are not met.

Propensity score analyses. While the teacher effect estimates from the various propensity score approaches had high correlations both with each other and with the value-added estimates (Tables 6, 13, and 14), these correlations were noticeably lower than within the sets of value-added estimates. Moreover, compared with the value-added estimates, the propensity score-based estimates were much more sensitive to variable-inclusion decisions. This sensitivity reflects the different use of the covariates. In the matching methods, a teacher's students were compared with the most similar students in the larger population. If the distribution of covariates was similar in the teacher's comparison group and the larger population, then the propensity score-based effect estimates were comparable with the value-added estimates, as shown when the matching ratio was increased. The increased correlation between the propensity score-based estimates and the value-added estimates as the matching ratio increased provided evidence of the non-comparability of the teachers' students to the population as a whole. When a teacher's students were matched to the best matches, the teacher's effect was found to differ from the value-added estimates. When the teacher's students were matched to an increasing pool of less-well matched students, then the teacher's effect estimate more closely matched the value-added estimate.

The propensity score estimation procedure also suggested that some teachers have classroom characteristic distributions so different from other classrooms that a balanced control group of similar students could not be found if too many matching variables were used. The number of teachers affected by this problem varied with the number of matching variables used. Essentially, if more matching variables were used, so that the criteria for matching become stricter, more teachers' classes became unmatchable. This result suggests both the need for more research to determine which variables are most essential for matching, and the need to

acknowledge that some teachers are working within contexts so unique that comparing their students' outcomes with those of other teachers is unjustifiable.

PSA offers the advantage of being able to address context-specific teacher effectiveness with a single number while still acknowledging differential teacher effectiveness across student types. The comparability is limited by the degree to which it is appropriate to compare how far a teacher's test scores diverged from the mean for their student distribution to how another teacher's test scores diverged from the mean for another student distribution. More work may need to be done to refine these comparisons, possibly at a secondary level. However, at minimum, it should be possible to identify which teachers are doing better or worse than expected for their student distributions.

The biggest problems with the propensity score approach come with the difficulty of covariate selection. Both regression models and propensity score approaches require the inclusion of all variables that relate to treatment selection and outcomes to be included in the model (the ignore treatment assignment assumption). However, unlike with linear models, loglinear models may become unestimable, or result in fitted values of zero or one, if selected covariates are not distributed across treatment and control groups. In both VA and PSA estimations, unequal covariate distributions can create bias in effect estimation. When estimating logistic models, however, any imbalance in covariates becomes obvious. This fact is both an advantage and a disadvantage. With PSA, it is impossible to proceed and estimate an ATE anyway, without making adjustments to the model. In complex situations with a large number of covariates, making the decisions necessary to achieve good model fit while still including any necessary covariates can be time-consuming and frustrating. It may become clear, for example, that an ATT, or no effect at all, can justifiably be estimated, when the desire is to

estimate the ATE. When fitting regression models, however, there is no such warning system. A researcher has much greater freedom to use covariates without limitations, resulting in greater room for misinterpretation of results.

Generalized propensity score analyses. The most important practical advantage of using multinomial models was that the propensity scores were estimated jointly, rather than independently, thus forcing them to sum to one across individuals and accounting for the clustering inherent in the pool from which potential controls were drawn. However, the estimated propensity scores using this approach were virtually identical ($r > .999$) to those estimated from the independent logit models, which had fewer convergence problems. Given that the multinomial models could not be estimated at all if too many groups or too many matching variables were modeled, the independent approach appears to be more practically useful.

A limitation of the present study was that each multinomial model included six or seven groups. The 435 teacher effect estimates were not, therefore, really estimated simultaneously. They were estimated in groups. However, experiments with fewer and greater numbers of groups, using the subsets of matching variables that made those various-sized models estimable, suggested that the generalized propensity scores would always be highly correlated with those from the independent models. This result is not unexpected, as no matter the number of equations estimated simultaneously, the counterfactual was identical. The propensity score always represented the probability of assignment to teacher A versus assignment to any teacher except teacher A.

At the effect estimation stage, Imbens' (2000) approach diverged dramatically from the matching method used in the independent propensity score analyses. His expected value

approach theoretically produces unbiased estimates of the average treatment effect (ATE), provided the ignorable treatment assignment assumption is met and the multinomial logit model is of correct form.

The lack of correlation of this set of teacher effects with any other set estimated in this study was a concern. However, it is important to note that the estimation of the ATE, as found using Imbens approach, can be dramatically different than the estimation of the average treatment effect for the treated (ATT), as found from matching-based effect estimates, especially in a large population. The first indicates how well a teacher would perform with any student in the population, while the second indicates how well a teacher performed with his actual assigned students. The more the teacher's assigned students differed from those within the population as a whole, the more the ATE and the ATT had potential to differ from each other. In essence, the lack of correlation of the expected value-based multinomial effects with the independent propensity score-based matched effects is further evidence in itself that teaching context matters, that unique classroom compositions can have dramatic impact on classroom-level outcomes, and that teacher evaluation and accountability systems need to exercise extreme care in making teacher comparisons.

It is also important to note that the value-added estimates were highly correlated with the matching-based ATT teacher effect estimates but not with the expected value-based ATE effect estimates derived from the multinomial models. While value-added estimates, like the expected value-based estimates, used all the data in one model, there is a significant difference. Essentially, value-added estimates are slopes on teacher dummy variables. The expected value estimates are predicted values from similar models, weighted according to the student's propensity to be in a given classroom. The value-added approach estimates a flat, constant

teacher effect across all students, given the covariates. The expected-value approach estimates a teacher effect that is allowed to vary by student, and weighted across the covariate distribution.

The most significant limitation of estimating the reliability of these expected value-based estimates of the ATE resulting from the multinomial logit models was that no other estimation method used in this study satisfied all assumptions necessary for ATE estimation. There was no gold standard to which the estimates could be compared. Because the multinomial models could not be estimated with more than 17 variables with the chosen strata size, the estimates likely suffered from violations of the ignorable treatment assignment assumption. In addition, the assumption could not be tested that the independent propensity score estimates would have equaled the multinomial generalized propensity scores if a simultaneous model with all 435 groups could have been estimated. If not, then using all sets of generalized propensity scores together in one joint estimation of expected values, as was done in this study, resulted in bias. It is possible that larger models were not estimatible simply because of the lack of coverage of covariates across groups, suggesting un-comparability of these groups. Further study with larger datasets may shed greater light on this issue.

Summary. The question of which set of estimates is least biased depends on both what assumptions we are willing to trust and what it is we wish to estimate. If we believe that teacher effects are fixed across all covariate levels, for example, then value-added estimates are relatively easy to find and will be unbiased, regardless of covariate distributions, provided all important covariates are modeled. Moreover, if teacher effects are fixed, distinctions between the ATE and the ATT become irrelevant because the teacher has the same effect on everyone—whether assigned to the classroom or belonging to the larger population—again, so long as all

relevant covariates are modeled. If teacher effects are truly fixed, and all relationships are linear, then there is no need to advance beyond using the standard linear regression model.

However, if teacher effects are not fixed—if specific teachers excel more with one student type than with another—then effect estimation becomes more problematic. Either differential effects must be modeled (Jakubowki, 2008; Lockwood & McCaffrey, 2009; Reardon & Raudenbush, 2009), or the teacher effect must be averaged across some distribution. Averaging the teacher effect across the entire population of interest (to estimate the ATE) requires that the teacher's class contain an adequate distribution of relevant characteristics found in that population. Even relatively large K-12 class sizes of 30-40 are too small for this requirement ever to be met across more than a fraction of classes. This problem leaves two remaining possibilities—average the effect across the teacher's own distribution of students (the ATT) or do not estimate the effect at all.

While the ATT, the average effect of the teacher on the teacher's own students, may seem limiting, it might be argued that this is exactly what we want. Teachers often specialize in teaching specific student types, and principals often become adept at matching students to teachers. It seems that a teacher should be rewarded for teaching well the assigned students, rather than some theoretical mix of students the teacher never encountered. While context-dependent ATTs may lose something in terms of comparability, they may more accurately reflect what we want teachers to do: that is, to teach well the students they are assigned.

Previous research addressing the stability of teacher effects across years suggests that teacher effect stability and bias (in the ATE) have an inverse relationship. When consistent student sorting occurs, such as when teachers become specialists, effect estimates are both more stable and more biased (McCaffrey et al., 2009). Teacher effects that reflect similar context

across years differ from the ATE, but do so consistently. Essentially, to the degree that a teacher specializes or has an unusual classroom composition, the effect being estimated is the ATT.

While PSA methods cannot account for unmeasured factors any better than VA models, they do acknowledge more fully the measurable covariates and the uniqueness of context. The limitations of disparate covariate distributions are more thoroughly recognized and dealt with using PSA than with VA modeling. Using PSA methods, it is more clear which effect, the ATE or ATT, is actually being estimated, and how strong the region of common support and matching quality is. In addition, sensitivity methods have been developed for propensity score-type matched effects which allow analysts to understand how sensitive the estimates are to unmeasured covariates (Guo & Fraser, 2010; Rosenbaum, 2002). Thus, in many respects, effects resulting from matching methods are more transparent than those from regression models.

Teachers should be empowered to teach their actual students and rewarded for doing that well. No statistical methodology yet offers a way to fairly compare all teachers with each other, taking into account the relative advantages and disadvantages of teaching context and student characteristics. The best that can be done statistically, at present, is to estimate how well a teacher does with the assigned students, relative to how other teachers do with similar students, given the characteristics that can actually be measured. Propensity score-based methods offer alternatives to regression-based methods that can complete that task more validly.

Further Research

Estimating the ATT rather than the ATE for teacher accountability appears to be a more obtainable goal from a statistical perspective. In fact, if student covariate distributions are inconsistent across classrooms, and if those inconsistencies matter—meaning individual teacher effect estimates depend on those covariate distributions, then estimating the ATE is beyond reach

using either linear models or matching-based methods. In spite of this problem, existing attempts at outcomes-based teacher evaluation, including value-added modeling, have estimating the ATE as the goal. These approaches endeavor to compare each teacher with all others within a system. It appears that the most valid estimates of teacher quality using student outcomes, however, will be made when comparisons are more limited.

To whom an individual teacher should be compared is, to some degree, a philosophical question that needs to be answered by policy makers, educators, and other stake holders. One area of further research that needs to be addressed is the potential for obtaining policy maker support for shifting the emphasis from teacher-compared-with-all to teacher-compared-with-like. In other words, is it possible to gain support for estimating the ATT as the correct effect of interest in teacher accountability efforts?

If so, then the focus of student achievement-based teacher effect estimation needs to shift. Rather than value-added linear models, matching methods or other approaches appropriate for estimating the ATT need to be more fully studied and developed. Variables that are both unbalanced across classrooms and predictive of student outcomes need to be identified and methods of measuring them reliably need to be developed. The practicality of measuring these variables within an ongoing accountability system needs to be evaluated and, where limitations are found, solutions need to be found. These solutions may include the development of sensitivity methods that can suggest the magnitude of the impact on teacher effect estimates of the inability to measure important variables. Finally, the impact on teachers, teaching, and learning of high stakes outcomes, based on the ATT teacher effect, needs to be examined.

Primum non nocere.

References

Allison, P.D. (1999). *Multiple regression: A primer*. Thousand Oaks, C.A.: Pine Forge Press.

Arpino, B., & Mealli, F. (2008). The specification of the propensity score in multilevel observational studies (Working Paper No. 6). Carlo F. Dondena Centre for Research on Social Dynamics. Retrieved from http:// www.dondena.unibocconi.it/wp6

Austin, P.C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46,* 399-424.

Baker, E., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., & Linn, R. L. (2010). Problems with the use of student test scores to evaluate teachers (Briefing Paper #278). Economic Policy Institute. Retrieved from http://epi.3cdn.net/b9667271ee6c154195_t9m6iij8k.pdf

Cochran, W. G., & Rubin, D. B. (1973). Controlling Bias in Observational Studies : A Review. *Sankhya: The Indian Journal of Statistics*, *35*(4), 417-446.

Caliendo, M., & Kopeinig, S. (2005). Some practical guidance for the implementation of propensity score matching (IZA DP No. 1588). Institute for the Study of Labor. Retrieved from http://ftp.iza.org/dp1588.pdf

Dattalo, P. (2010). *Strategies to approximate random sampling and assignment.* New York: Oxford University Press.

Doyle, W. R. (2011). Effect of increased academic momentum on transfer rates: An application of the generalized propensity score. *Economics of Education Review*, *30*(1), 191-200. Elsevier Ltd. doi: 10.1016/j.econedurev.2010.08.004.

Eckert, J.M., & Dabrowski, J. (2010). Should value-added measures be used for performance pay? *Phi Delta Kappan*, *91*(8), 88-92.

Ertefaie, A., & Stephens, D. (2010). Comparing approaches to causal inference for longitudinal data: Inverse probability weighting versus propensity scores. *The International Journal of Biostatistics*, *6*(2). doi: 10.2202/1557-4679.1198.

Foster, E.M. (2003). Propensity score matching: An illustrative analysis of dose response. *Medical Care*, *41*(10), 1183-1192.

Fryges, H. (2009). The export–growth relationship: Estimating a dose-response function. *Applied Economics Letters*, *16*(18), 1855-1859. doi: 10.1080/13504850701719496.

Gingerich, D.W. (2010). Understanding off-the-books politics: Conducting inference on the determinants of sensitive behavior with randomized response surveys. *Political Analysis*, *18*(3), 349-380. doi: 10.1093/pan/mpq010.

Goldhaber, D., & Hansen, M. (2008). Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions (Brief 3). National Center for Analysis of Longitudinal Data in Education Research. Retrieved from http://www.caldercenter.org/upload/Teacher_Job_Performance.pdf

Griswold, M.E., & Localio, A.R. (2010). Propensity score adjustment with multilevel data: Setting your sites on decreasing selection bias. *Annals of Internal Medicine, 152*(6), 393-396.

Guo, G., & Zhao, H. (2000). Multilevel modeling for binary data. *Annual Review of Sociology*, *26*(1), 441-462. doi: 10.1146/annurev.soc.26.1.441.

Guo, S., & Fraser, M.W. (2010). *Propensity score analysis*. Los Angeles: SAGE.

Hahs-Vaughn, D.L., & Onwuegbuzie, A.J. (2006). Estimating and using propensity score analysis with complex samples. *The Journal of Experimental Education*, *75*(1), 31-65. doi: 10.3200/JEXE.75.1.31-65.

Harris, D.N. (2009). Teacher value-added: Don't end the search before it starts. *Journal of Policy Analysis and Management*, *28*(4), 693-699.

Harris, D.N., & McCaffrey, D.F. (2010). Value-added: Assessing teachers' contributions to student achievement. In M. Kennedy (Ed.), *Teacher Assessment and the Quest for Teacher Quality* (pp. 251-282). San Francisco: Jossey-Bass.

Hodgens, J. L., & Lehmann, E. L. (1962). Rank methods for combination of independent experiments in analysis of variance. *The Annals of Mathematical Statistics*, *33*(2), 482-497.

Hong, G. (2010). Marginal mean weighting through stratification: Adjustment for selection bias in multilevel data. *Journal of Educational and Behavioral Statistics*, *35*(5), 499-531. doi: 10.3102/1076998609359785.

Hong, G., & Raudenbush, S.W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis*, *27*(3), 205-224. doi: 10.3102/01623737027003205.

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

Hong, Guanglei, & Yu, Bing. (2008). Effects of kindergarten retention on children's social-emotional development: An application of propensity score method to multivariate, multilevel data. *Developmental Psychology*, *44*(2), 407-21. doi: 10.1037/0012-1649.44.2.407.

Imai, K., & van Dyk, D. (2004). Causal inference with general treatment regimes. *Journal of the American Statistical Association*, *99*(467), 854-866. doi: 10.1198/016214504000001187.

Imbens, G.W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*(3), 706-710.

Imbens, G.W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics, 86*(1), 4-29.

Ishii, J., & Rivkin, S.G. (2009). Impediments to the estimation of teacher value-added. *Education Finance and Policy*, *4*(4), 520-536.

Jakubowski, M. (2008). Implementing value-added models of school assessment (EUI RSCAS Working Paper 2008/06). European University Institute. Retrieved from http://ideas.repec.org/p/rsc/rsceui/2008-06.html

Joffe, M.M., & Rosenbaum, Paul R. (1999). Invited commentary: Propensity scores. *American Journal of Epidemiology*, *155*(8), 328-333. doi: 10.1002/ajmg.a.34236.

Kleinbaum, D.G., & Klein, M. (2002). *Logistic regression: A self-learning text*. New York: Springer.

Koedel, C., & Betts, J.R. (2007). Re-examining the role of teacher quality in the educational production function (Working Papers Series WP 09-02). University of Missouri. Retrieved from http://economics.missouri.edu/working-papers/2007/wp0708_koedel.pdf

Koedel, C., & Betts, J.R. (2009). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique (Working Paper 2009-01). *National Center on Performance Incentives*. Retrieved from http://economics.missouri.edu/working-papers/2009/WP0902_koedel.pdf

Lambert, J. (2008, September 12). Texas teacher wins lawsuit on wrongful termination. *Education Week Teacher*. Retrieved from http://blogs.edweek.org/teachers/webwatch/2008/09/texas_teacher_wins_lawsuit_on.html

Lechner, M. (1999). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption (Discussion Paper No. 91). IZA. Retrieved from ftp://repec.iza.org/RePEc/Discussionpaper/dp91.pdf

Leon County School Board v. Waters, 2007 WL 200601 (1986).

Levine, D. I., & Painter, G. (2008). Are measured school effects just sorting? *Economics of Education Review*, *27*(4), 460-470.

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, *45*(1), 255-68. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2720055.

Lockwood, J.R., & McCaffrey, D.F. (2009). Exploring student-teacher interactions in longitudinal achievement data. *Education Finance and Policy*, *4*(4), 439-467.

Lockwood, J.R., McCaffrey, D.F., Hamilton, L.S., Stecher, B., Le, V-N., & Martinez, J.F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, *44*(1), 47-67.

Lu, B., Zanutto, E., Hornik, R., & Rosenbaum,P. (2011). Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association*, *96*(456), 1245- 1253.

Massachusetts Federation of Teachers, AFT, AFL-CIO v. Board of Education, 436 Mass. 863 (2002).

McCaffrey, D.F., Sass, T.R., Lockwood, J.R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy*, *4*(4), 572-606.

Morgan, S.L. (2001). Counterfactuals, causal effect heterogeneity, and the catholic school effect on learning. *Sociology of Education*, *74*(4), 341. doi: 10.2307/2673139.

Morgan, S.L. & Winship, C. (2007). *Counterfactuals and causal inference*. New York: Cambridge University Press.

Mulrow, C. (2010). Propensity score adjustment with multilevel data: Setting your sites on decreasing selection bias. *Annals of Internal Medicine*, *152*(6), 393-396.

National Conference of State Legislatures (2010, November 16). *Education Bill Tracking Database.* Retrieved from http://www.ncsl.org/IssuesResearch/Education/EducationBillTrackingDatabase/tabid/12913/Default.aspx

Newton, X.A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives*, *18*(23), 1-24.

R Development Core Team (2011). R: A language and environment for statistical computing (Version 2.14.1) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Program available at http://www.r-project.org

Reardon, S.F., & Raudenbush, S.W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy*, *4*(4), 492-519.

Rock, D.A., & Pollack, J.M. (2002). Early childhood longitudinal study--kindergarten class of 1998-99 (ECLS-K): Psychometric report for kindergarten through first grade (Working Paper NCES-WP-2002-05). National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubsearch

Rosenbaum, P. (2002). *Observational Studies* (2nd ed.). New York: Springer.

Rosenbaum, Paul R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41. doi: 10.2307/2335942.

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

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*, 688-701.

Rubin, D.B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology, 2*, 169-188.

Rubin, D.B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association, 100,* 322-331.

Sass, T.R. (2008). The stability of value-added measures of teacher quality and implications for teacher compensation policy (Brief 4). National Center for Analysis of Longitudinal Data in Education Research. Retrieved from http://www.urban.org/publications/1001266.html

Scheelhaase v. Woodbury Central Community School District, 488 F. 2d 237 (1973).

Schochet, P.Z., & Chiang, H.S. (2010). Error rates in measuring teacher and school performance based on student test score gains (NCEE 2010-4004). Institute of Education Sciences. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED511026

Schreyogg, J., Stargardt, T., & Tiemann, O. (2011). Costs and quality of hospitals in different health care systems: A multi-level approach with propensity score matching. *Health Economics*, *100*(July 2010), 85-100. doi: 10.1002/hec.

Sherrod v. Palm Beach County School Board, 963 So. 2d 251, (2006).

Smyth, E. (2008). The more, the better? Intensity of involvement in private tuition and examination performance. *Educational Research and Evaluation*, *14*(5), 465-476. doi: 10.1080/13803610802246395.

St. Louis Teachers Union, Local 420, American Federation of Teachers, AFL-CIO v. Board of Education of the City of St. Louis, 652 F. Supp. 425 (1987).

Stuart, E. a. (2007). Estimating causal effects using school-level data sets. *Educational Researcher*, *36*(4), 187-198. doi: 10.3102/0013189X07303396.

Thoemmes, F.J., & Kim, E.S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research*, *46*(1), 90-118.

Thoemmes, F.J., & West, S.G. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research*, *46*(3), 514-543.

Tourangeau, K., Nord, C., Le, T., Sorongon, A. G., Najarian, M., & Hausken, E.G. (2009). *Early childhood longitudinal study, kindergarten class of 1998-99 (ECLS-K): Combined user's manual for the ECLS-K eighth-grade and K-8 full sample data files and electronic codebooks*. National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubsearch

Winship, C. & Morgan, S.L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology, 25*, 659-706.

Young v. Palm Beach County School Board, 968 So. 2d 38 (2006).

Zanutto, E., Lu, B., & Hornik, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics*, *30*(1), 59-73. doi: 10.3102/10769986030001059.

Appendix A: Comparability of Teacher Control Groups

One point of complexity in the propensity score analyses using the logit model is that teacher effects will need to be comparable to each other. Because each teacher's control state is slightly different than another's (not in teacher *j*'s class versus not in teacher *m*'s class), this raises the issue of comparability.

Assume r_{1ij} represents the theoretical outcome for student *i* in class *j*, and r_{0ij} represents the theoretical outcome for student *i* in *not* class *j* (notation adapted from Rosenbaum & Rubin, 1983). We might define the ATE for teacher *j* as $\alpha_j = E(r_{1j}) - E(r_{0j})$, and for treatment or teacher *m* as $\alpha_m = E(r_{1m}) - E(r_{0m})$, where $j \neq m$. In each case, 1 represents the teacher's treatment state (taught by that teacher) and 0 represents the teacher's control state (not taught by that teacher). The difference between the true treatment effects for teachers *j* and *m* is:

$$\alpha_{j,m} = \alpha_j - \alpha_m = (E(r_{1j}) - E(r_{0j})) - (E(r_{1m}) - E(r_{0m})). \quad (3)$$

If $E(r_{0j}) = E(r_{0m})$, then the difference in treatment effects for teachers *j* and *m* simplifies to:

$$\alpha_{j,m} = E(r_{1j}) - E(r_{1m}). \quad (4)$$

Essentially, the mean treatment effects for the two teachers can be directly compared if the two control *states* are identical—or more precisely, the expected values of the outcome in each of the two control states are equivalent.

In this sampling design, the control state for teacher *j* is "taught by any teacher but *j*". The control state for teacher *m* is "taught by any teacher but *m*". If there are $k = 2$ teachers in the study, then the control state for each teacher is the other teacher's treatment state, and there is no overlap between the two control states.

However, it is clear that as *k* increases, the overlap in the control states increases across teachers, and the differences in the control states becomes trivial, reducing the bias due to the

shifting control groups. It is the large number of teacher/treatment effects that are typically estimated in state-wide accountability systems that make this approach feasible. In other words, k is typically very large in practical settings, when teachers are compared across schools. Therefore, across-school comparisons of teachers should be comparable.

Appendix B: Original List of 187 Potential Covariates

Variable Name	Variable Label
C1ASMTDD	C1 ASSESSMENT DAY
C1ASMTMM	C1 ASSESSMENT MONTH
C1ASMTST	C1 CHILD ASSESSMENT STATUS
C1ASMTYY	C1 ASSESSMENT YEAR
C1BMI	C1 ROUND 1 CHILD COMPOSITE BMI
C1CMOTOR	C1 COMPOSITE MOTOR SKILLS
C1FMOTOR	C1 FINE MOTOR SKILLS
C1GMOTOR	C1 GROSS MOTOR SKILLS
C1HEIGHT	C1 ROUND 1 CHILD COMPOSITE HGT (INCHES)
C1R4MPB1	C1 RC4 PROB1 - COUNT, NUMBER, SHAPE
C1R4MPB2	C1 RC4 PROB2 - RELATIVE SIZE
C1R4MPB3	C1 RC4 PROB3 - ORDINALITY, SEQUENCE
C1R4MPB4	C1 RC4 PROB4 - ADD/SUBTRACT
C1R4MPB5	C1 RC4 PROB5 - MULTIPLY/DIVIDE
C1R4MPB6	C1 RC4 PROB6 - PLACE VALUE
C1R4MPB7	C1 RC4 PROB7 - RATE & MEASUREMENT
C1R4MPB8	C1 RC4 PROB8 - FRACTIONS
C1R4MPB9	C1 RC4 PROB9 - AREA AND VOLUME
C1R4MSCL	C1 RC4 MATH IRT SCALE SCORE
C1R4RP10	C1 RC4 PROB10 - EVALUATE COMPLEX SYNTAX
C1R4RPB1	C1 RC4 PROB1 - LETTER RECOGNITION
C1R4RPB2	C1 RC4 PROB2 - BEGINNING SOUNDS
C1R4RPB3	C1 RC4 PROB3 - ENDING SOUNDS
C1R4RPB4	C1 RC4 PROB4 - SIGHT WORDS
C1R4RPB5	C1 RC4 PROB5 - WORD IN CONTEXT
C1R4RPB6	C1 RC4 PROB6 - LITERAL INFERENCE
C1R4RPB7	C1 RC4 PROB7 - EXTRAPOLATION
C1R4RPB8	C1 RC4 PROB8 - EVALUATION
C1R4RPB9	C1 RC4 PROB9 - EVALUATING NON-FICTION
C1R4RSCL	C1 RC4 READING IRT SCALE SCORE
C1RGSCAL	C1 REC GENERAL KNOWLEDGE IRT SCALE SCORE
C1RRPRIN	C1 PRINT FAMILIARITY
C1SCREEN	C1 SPEAK NON-ENGLISH LANGUAGE AT HOME
C1SCSTO	C1 AIQ400 TELL STORIES CHILD SCORE
C1SCTOT	C1 AIQ400 CHILD'S TOTAL OLDS SCORE
C1SPASMT	C1 CHILD ASSESSMENT IN SPANISH

(Appendix B continues)

(Appendix B continued)

Variable Name	Variable Label
C1SPHOME	C1 SPEAK SPANISH AT HOME
C1SSCART	C1 SAI400 SPANISH ART SHOW CHILD SCORE
C1SSCORD	C1 SAI400 SPANISH SIMON SAYS CHILD SCORE
C1SSCSTO	C1 SAI400 SPANISH TELL STORIES SCORE
C1SSCTOT	C1 SAI400 SPANISH TOTAL OLDS SCORE
C1WEIGHT	C1 ROUND 1 CHILD COMPOSITE WGT (POUNDS)
F2SPECS	F2 CHILD RECEIVED SPEC EDU SERV FROM FMS
GENDER	CHILD COMPOSITE GENDER
P1ACTIV2	P1 CHQ145 CONCERNS - CHD ACTIVITY LEVEL
P1ADLTLV	P1 HRQ130 ADULTS LIVING WITH CHILD
P1AGEENT	P1 AGE (MONTHS) AT KINDERGARTEN ENTRY
P1AGEFRS	P1 AGE (MNTHS) AT FIRST NONPARENTAL CARE
P1ANYLNG	P1 PLQ020 IF OTHER LANGUAGE USED AT HOME
P1BDAGE	P1 AGE OF NONRES BIO FATHER (YRS)
P1BDMT1R	P1 NONRES BIO FATHER MORE THAN 1 RACE
P1BDRACE	P1 RACE OF NONRES BIOLOGICAL FATHER
P1BEHAVE	P1 CHQ325 BEHAVES AS WELL AS OTHER CHDN
P1BMAFB	P1 AGE AT 1ST BIRTH NONRES BIO MOM (YRS)
P1BMAGE	P1 AGE OF NONRES BIO MOTHER (YRS)
P1BMMT1R	P1 NONRES BIO MOTHER MORE THAN 1 RACE
P1BMRACE	P1 RACE OF NONRES BIOLOGICAL MOTHER
P1BUILD	P1 HEQ010 HOW OFTEN YOU ALL BUILD THINGS
P1CARNOW	P1 CURRENT NONPARENTAL CARE ARRANGEMENTS
P1CENTER	P1 CHILD EVER IN CENTER-BASED CARE
P1CHLAUD	P1 HEQ050 HOW MANY RECORDS, TAPES, CDS
P1CHLBOO	P1 HEQ040 HOW MANY BOOKS CHILD HAS
P1CHLPIC	P1 HEQ060 HOW OFTEN READS PICTURE BOOKS
P1CHOOSE	P1 PIQ050 CURR SCHOOL AFFECT HOME CHOICE
P1CHORES	P1 HEQ010 HOW OFTEN CHILD DOES CHORES
P1CHREAD	P1 HEQ070 FREQ READS BOOKS OUTSIDE SCH
P1CHSESA	P1 HEQ080 PRE K CHILD WATCHED SESAME ST
P1COMPLI	P1 CHQ085 OTHER BIRTH COMPLICATIONS
P1CONTRO	P1 SELF-CONTROL
P1DADOCC	P1 RESIDENT FATHER'S OCCUPATION
P1DIAGNO	P1 CHQ120 LEARNING PROBLEM DIAGNOSED
P1DIFFHR	P1 CHQ230 IF DIFFICULTY HEARING SPEECH
P1DISABL	P1 CHILD W/ DISABILITY

(Appendix B continues)

(Appendix B continued)

Variable Name	Variable Label
P1EARIN2	P1 CHQ327 IF CHD OFTEN HAD EAR INFECTION
P1EARINF	P1 CHQ326 IF CHD OFTEN HAS EAR INFECTION
P1EARLY	P1 CHQ030 HOW PREMATURE - NUMBER
P1ENGLIS	P1 PLQ030 IF ENGLISH ALSO USED AT HOME
P1EVALUA	P1 CHQ115 CHD LEARNING ABILITY EVALUATED
P1FIRKDG	P1 FIRST-TIME KINDERGARTENER
P1FTHGRD	P1 PEQ140 RESP FATHER HIGHEST ED LEVEL
P1GAMES	P1 HEQ010 HOW OFTEN YOU ALL PLAY GAMES
P1HDAD	P1 RESIDENT FATHER TYPE
P1HDAGE	P1 AGE - CURRENT FATHER (YRS)
P1HDEMP	P1 CURRENT FATHER EMPLOYMENT STATUS
P1HDLANG	P1 FATHER'S LANGUAGE TO CHILD
P1HDLTOD	P1 CHILD'S LANGUAGE TO FATHER
P1HDMT1R	P1 FATHER MORE THAN ONE RACE
P1HDRACE	P1 RACE OF CURRENT FATHER
P1HELPAR	P1 HEQ010 HOW OFTEN YOU HELP CHD DO ART
P1HFAMIL	P1 FAMILY TYPE
P1HIG_1	P1 PEQ020 PERS 1 HIGHEST EDUCATION LEVEL
P1HIG_2	P1 PEQ020 PERS 2 HIGHEST EDUCATION LEVEL
P1HIGHSC	P1 PEQ100 RESP'S GRADES IN HIGH SCHOOL
P1HIS_1	P1 PEQ030 IF PERS 1 HIGH SCHOOL DIPLOMA
P1HIS_2	P1 PEQ030 IF PERS 2 HIGH SCHOOL DIPLOMA
P1HMAFB	P1 AGE AT 1ST BIRTH - CURRENT MOM (YRS)
P1HMAGE	P1 AGE - CURRENT MOTHER (YRS)
P1HMEMP	P1 CURRENT MOTHER EMPLOYMENT STATUS
P1HMLANG	P1 MOTHER'S LANGUAGE TO CHILD
P1HMLTOM	P1 CHILD'S LANGUAGE TO MOTHER
P1HMMT1R	P1 MOTHER MORE THAN ONE RACE
P1HMOM	P1 RESIDENT MOTHER TYPE
P1HMRACE	P1 RACE OF CURRENT MOTHER
P1HOWOLD	P1 CHQ130 AGE AT 1ST DIAGNS-LRN ABLTY
P1HRSNOW	P1 # HOURS SPENT IN NONPARENTAL CARE NOW
P1HRSPRK	P1 # HRS SPENT IN NONPARENTAL CARE PRE-K
P1HSCALE	P1 CHQ330 1-5 SCALE OF CHILD'S HEALTH
P1HSDAYS	P1 CCQ250 # OF DAYS/WK IN HEAD START
P1HSEVER	P1 CCQ210 WAS CHILD EVER IN HEAD START

(Appendix B continues)

(Appendix B continued)

Variable Name	Variable Label
P1HSHRS	P1 CCQ251 # OF HRS/WK IN HEAD START
P1HSPREK	P1 CCQ215 IN HS YEAR BEFORE K
P1HSTYPE	P1 CCQ245 IN HEAD START FULL OR PART-DAY
P1HTOTAL	P1 TOTAL NUMBER IN HOUSEHOLD
P1IMPULS	P1 IMPULSIVE/OVERACTIVE
P1LANGUG	P1 BOTH PARENT LANGUAGE TO CHILD
P1LEARN	P1 APPROACHES TO LEARNING
P1LEGMAR	P1 MHQ020 RESBIODAD MARRIED TO RESBIOMOM
P1LESS18	P1 NUMBER IN HOUSEHOLD AGED <18
P1MMDIAG	P1 CHQ135 MNTH AT 1ST DIAGNS-LRN ABLTY
P1MOMOCC	P1 RESIDENT MOTHER'S OCCUPATION
P1MTEACH	P1 PIQ030 HAVE YOU MET CHILD'S TEACHER
P1MTHGRD	P1 PEQ150 RESP MOTHER HIGHEST ED LEVEL
P1MULTIP	P1 CHQ035 CHILD PART OF MULTIPLE BIRTH
P1NATURE	P1 HEQ010 HOW OFTEN YOU TEACH CHD NATURE
P1NUMARR	P1 CCQ030 # REL CARE ARRANGE YR BEFORE K
P1NUMNOW	P1 # NONPARENTAL CARE ARRANGEMENTS NOW
P1NUMSIB	P1 NUMBER OF SIBLINGS IN HOUSEHOLD
P1OVER18	P1 NUMBER IN HOUSEHOLD AGED 18+
P1PREMAT	P1 CHQ025 MORE THAN 2 WEEKS EARLY
P1PRIMNW	P1 PRIMARY TYPE OF NONPARENTAL CARE
P1PRIMPK	P1 PRIMARY TYPE NONPARENTAL CARE PRE-K
P1PRMLNG	P1 PLQ060 WHAT PRIMARY LANGUAGE AT HOME
P1PRONO2	P1 CHQ205 IF CHILD HAD SPEECH PROBLEMS
P1RCHLD	P1 CCQ095 # CHILDREN CARED FOR TOGETHER
P1RDAYPK	P1 CCQ040 # DAYS/WK REL CARE YR BEFORE K
P1RDAYS	P1 CCQ085 # OF DAYS/WK OF REL CARE
P1READBO	P1 HEQ010 HOW OFTEN YOU READ TO CHILD
P1READEN	P1 PLQ070 HOW WELL RESP READ ENGLISH
P1RHRS	P1 CCQ090 # OF HRS/WK OF REL CARE
P1RHRSPK	P1 CCQ045 # HRS/WK REL CARE YR BEFORE K
P1RMOPK	P1 CCQ050 # MONTHS REL CARE YR BEFORE K
P1SADLON	P1 SAD/LONELY
P1SIGHT	P1 CHQ285 DIFFICULT SEEING FAR OBJECT
P1SINGSO	P1 HEQ010 HOW OFTEN YOU ALL SING SONGS
P1SOCIAL	P1 SOCIAL INTERACTION
P1SPEAKE	P1 PLQ070 HOW WELL RESP SPEAK ENGLISH

(Appendix B continues)

(Appendix B continued)

Variable Name	Variable Label
P1SPORT	P1 HEQ010 HOW OFTEN YOU ALL DO SPORTS
P1STPREP	P1 PIQ020 CHILD KINDERGARTEN PREPARATION
P1TELLST	P1 HEQ010 HOW OFTEN YOU TELL CHD STORIES
P1THER10	P1 CHQ345 SPECIAL NEEDS CLASSES
P1THER11	P1 CHQ345 PRIVATE TUTORING
P1THERA2	P1 CHQ345 SPEECH THERAPY
P1THERA3	P1 CHQ345 OCCUPATIONAL THERAPY
P1THERA4	P1 CHQ345 PHYSICAL THERAPY
P1THERA5	P1 CHQ345 VISION SERVICES
P1THERA6	P1 CHQ345 SOCIAL WORK SERVICES
P1THERA7	P1 CHQ345 PSYCHOLOGICAL SERVICES
P1THERAP	P1 CHQ340 IF THERAPY BEFORE SCHOOL YEAR
P1TWINST	P1 CHILD BIRTH STATUS
P1UNDERS	P1 PLQ070 HOW WELL RESP UNDERSTAND ENG
P1WEIGH5	P1 CHQ010 MORE THAN 5.5 POUNDS AT BIRTH
P1WEIGH6	P1 CHQ015 MORE THAN 3 POUNDS AT BIRTH
P1WEIGHO	P1 CHQ005 CHILD WEIGHT AT BIRTH - OUNCES
P1WEIGHP	P1 CHQ005 CHILD WEIGHT AT BIRTH - POUNDS
P1WHATDI	P1 CHQ125 1ST DIAGNOSIS-LEARNING ABILITY
P1WHICHY	P1 PIQ080 CHILD'S YEAR OF KINDERGARTEN
P1WICCHD	P1 WPQ040 WIC BENEFITS FOR CHILD
P1WICMOM	P1 WPQ030 WIC BENEFITS WHEN PREGNANT
P1WRITEN	P1 PLQ070 HOW WELL RESP WRITE ENGLISH
P2FREERD	P2 WPQ180 FREE OR REDUCED LUNCH
P2HILOW	P2 PAQ110 INCOME- MORE/LESS THAN 25K
P2INCOME	P2 PAQ100 TOTAL HOUSEHOLD INCOME ($)
P2LUNCHS	P2 WPQ170 CHD RECVS FREE/RED PRICE LUNCH
R1_KAGE	R1 COMPOSITE CHILD ASSESSMENT AGE(MNTHS)
R2_KAGE	R2 COMPOSITE CHILD ASSESSMENT AGE(MNTHS)
RACE	CHILD COMPOSITE RACE
T1CONTRO	T1 SELF-CONTROL
T1EXTERN	T1 EXTERNALIZING PROBLEM BEHAVIORS
T1INTERN	T1 INTERNALIZING PROBLEM BEHAVIORS
T1INTERP	T1 INTERPERSONAL
T1LEARN	T1 APPROACHES TO LEARNING
T1RARSGE	T1 REC GENERAL KNOWLEDGE ARS SCORE
T1RARSLI	T1 REC LITERACY ARS SCORE

(Appendix B continues)

(Appendix B continued)

Variable Name	Variable Label
T1RARSMA	T1 REC MATH ARS SCORE
W1INCOME	W1 INCOME (IMPUTED)
W1SESL	W1 CONTINUOUS SES MEASURE
WKSESL	WK CONTINUOUS SES MEASURE

Appendix C: Reduced List of 111 Potential Covariates

Variable Name	Variable Label
C1ASMTMM	C1 ASSESSMENT MONTH
C1BMI	C1 ROUND 1 CHILD COMPOSITE BMI
C1CMOTOR	C1 COMPOSITE MOTOR SKILLS
C1FMOTOR	C1 FINE MOTOR SKILLS
C1GMOTOR	C1 GROSS MOTOR SKILLS
C1HEIGHT	C1 ROUND 1 CHILD COMPOSITE HGT (INCHES)
C1R4MPB1	C1 RC4 PROB1 - COUNT, NUMBER, SHAPE
C1R4MPB2	C1 RC4 PROB2 - RELATIVE SIZE
C1R4MPB3	C1 RC4 PROB3 - ORDINALITY, SEQUENCE
C1R4MSCL	C1 RC4 MATH IRT SCALE SCORE
C1R4RPB1	C1 RC4 PROB1 - LETTER RECOGNITION
C1R4RPB2	C1 RC4 PROB2 - BEGINNING SOUNDS
C1R4RPB3	C1 RC4 PROB3 - ENDING SOUNDS
C1R4RSCL	C1 RC4 READING IRT SCALE SCORE
C1RGSCAL	C1 REC GENERAL KNOWLEDGE IRT SCALE SCORE
C1RRPRIN	C1 PRINT FAMILIARITY
C1SCREEN	C1 SPEAK NON-ENGLISH LANGUAGE AT HOME
C1SPHOME	C1 SPEAK SPANISH AT HOME
C1WEIGHT	C1 ROUND 1 CHILD COMPOSITE WGT (POUNDS)
GENDER	CHILD COMPOSITE GENDER
P1ADLTLV	P1 HRQ130 ADULTS LIVING WITH CHILD
P1AGEENT	P1 AGE (MONTHS) AT KINDERGARTEN ENTRY
P1AGEFRS	P1 AGE (MNTHS) AT FIRST NONPARENTAL CARE
P1ANYLNG	P1 PLQ020 IF OTHER LANGUAGE USED AT HOME
P1BEHAVE	P1 CHQ325 BEHAVES AS WELL AS OTHER CHDN
P1BUILD	P1 HEQ010 HOW OFTEN YOU ALL BUILD THINGS
P1CARNOW	P1 CURRENT NONPARENTAL CARE ARRANGEMENTS
P1CENTER	P1 CHILD EVER IN CENTER-BASED CARE
P1CHLAUD	P1 HEQ050 HOW MANY RECORDS, TAPES, CDS
P1CHLBOO	P1 HEQ040 HOW MANY BOOKS CHILD HAS
P1CHLPIC	P1 HEQ060 HOW OFTEN READS PICTURE BOOKS
P1CHOOSE	P1 PIQ050 CURR SCHOOL AFFECT HOME CHOICE
P1CHORES	P1 HEQ010 HOW OFTEN CHILD DOES CHORES
P1CHREAD	P1 HEQ070 FREQ READS BOOKS OUTSIDE SCH
P1CHSESA	P1 HEQ080 PRE K CHILD WATCHED SESAME ST
P1COMPLI	P1 CHQ085 OTHER BIRTH COMPLICATIONS

(Appendix C continues)

(Appendix C continued)

Variable Name	Variable Label
P1CONTRO	P1 SELF-CONTROL
P1DADOCC	P1 RESIDENT FATHER'S OCCUPATION
P1DISABL	P1 CHILD W/ DISABILITY
P1EARIN2	P1 CHQ327 IF CHD OFTEN HAD EAR INFECTION
P1EARINF	P1 CHQ326 IF CHD OFTEN HAS EAR INFECTION
P1FTHGRD	P1 PEQ140 RESP FATHER HIGHEST ED LEVEL
P1GAMES	P1 HEQ010 HOW OFTEN YOU ALL PLAY GAMES
P1HDAD	P1 RESIDENT FATHER TYPE
P1HDAGE	P1 AGE - CURRENT FATHER (YRS)
P1HDEMP	P1 CURRENT FATHER EMPLOYMENT STATUS
P1HDLANG	P1 FATHER'S LANGUAGE TO CHILD
P1HDLTOD	P1 CHILD'S LANGUAGE TO FATHER
P1HDRACE	P1 RACE OF CURRENT FATHER
P1HELPAR	P1 HEQ010 HOW OFTEN YOU HELP CHD DO ART
P1HFAMIL	P1 FAMILY TYPE
P1HIGHSC	P1 PEQ100 RESP'S GRADES IN HIGH SCHOOL
P1HIS_1	P1 PEQ030 IF PERS 1 HIGH SCHOOL DIPLOMA
P1HMAFB	P1 AGE AT 1ST BIRTH - CURRENT MOM (YRS)
P1HMAGE	P1 AGE - CURRENT MOTHER (YRS)
P1HMEMP	P1 CURRENT MOTHER EMPLOYMENT STATUS
P1HMLANG	P1 MOTHER'S LANGUAGE TO CHILD
P1HMLTOM	P1 CHILD'S LANGUAGE TO MOTHER
P1HMOM	P1 RESIDENT MOTHER TYPE
P1HMRACE	P1 RACE OF CURRENT MOTHER
P1HRSNOW	P1 # HOURS SPENT IN NONPARENTAL CARE NOW
P1HRSPRK	P1 # HRS SPENT IN NONPARENTAL CARE PRE-K
P1HSCALE	P1 CHQ330 1-5 SCALE OF CHILD'S HEALTH
P1HSEVER	P1 CCQ210 WAS CHILD EVER IN HEAD START
P1HTOTAL	P1 TOTAL NUMBER IN HOUSEHOLD
P1IMPULS	P1 IMPULSIVE/OVERACTIVE
P1LANGUG	P1 BOTH PARENT LANGUAGE TO CHILD
P1LEARN	P1 APPROACHES TO LEARNING
P1LESS18	P1 NUMBER IN HOUSEHOLD AGED <18
P1MOMOCC	P1 RESIDENT MOTHER'S OCCUPATION
P1MTHGRD	P1 PEQ150 RESP MOTHER HIGHEST ED LEVEL
P1NATURE	P1 HEQ010 HOW OFTEN YOU TEACH CHD NATURE
P1NUMARR	P1 CCQ030 # REL CARE ARRANGE YR BEFORE K

(Appendix C continues)

(Appendix C continued)

Variable Name	Variable Label
P1NUMNOW	P1 # NONPARENTAL CARE ARRANGEMENTS NOW
P1NUMSIB	P1 NUMBER OF SIBLINGS IN HOUSEHOLD
P1OVER18	P1 NUMBER IN HOUSEHOLD AGED 18+
P1PREMAT	P1 CHQ025 MORE THAN 2 WEEKS EARLY
P1PRIMNW	P1 PRIMARY TYPE OF NONPARENTAL CARE
P1PRIMPK	P1 PRIMARY TYPE NONPARENTAL CARE PRE-K
P1PRONO2	P1 CHQ205 IF CHILD HAD SPEECH PROBLEMS
P1RDAYPK	P1 CCQ040 # DAYS/WK REL CARE YR BEFORE K
P1READBO	P1 HEQ010 HOW OFTEN YOU READ TO CHILD
P1RHRSPK	P1 CCQ045 # HRS/WK REL CARE YR BEFORE K
P1RMOPK	P1 CCQ050 # MONTHS REL CARE YR BEFORE K
P1SADLON	P1 SAD/LONELY
P1SINGSO	P1 HEQ010 HOW OFTEN YOU ALL SING SONGS
P1SOCIAL	P1 SOCIAL INTERACTION
P1SPORT	P1 HEQ010 HOW OFTEN YOU ALL DO SPORTS
P1STPREP	P1 PIQ020 CHILD KINDERGARTEN PREPARATION
P1TELLST	P1 HEQ010 HOW OFTEN YOU TELL CHD STORIES
P1WEIGH5	P1 CHQ010 MORE THAN 5.5 POUNDS AT BIRTH
P1WEIGHO	P1 CHQ005 CHILD WEIGHT AT BIRTH - OUNCES
P1WEIGHP	P1 CHQ005 CHILD WEIGHT AT BIRTH - POUNDS
P1WICCHD	P1 WPQ040 WIC BENEFITS FOR CHILD
P1WICMOM	P1 WPQ030 WIC BENEFITS WHEN PREGNANT
P2INCOME	P2 PAQ100 TOTAL HOUSEHOLD INCOME ($)
P2LUNCHS	P2 WPQ170 CHD RECVS FREE/RED PRICE LUNCH
R1_KAGE	R1 COMPOSITE CHILD ASSESSMENT AGE(MNTHS)
R2_KAGE	R2 COMPOSITE CHILD ASSESSMENT AGE(MNTHS)
RACE	CHILD COMPOSITE RACE
T1CONTRO	T1 SELF-CONTROL
T1EXTERN	T1 EXTERNALIZING PROBLEM BEHAVIORS
T1INTERN	T1 INTERNALIZING PROBLEM BEHAVIORS
T1INTERP	T1 INTERPERSONAL
T1LEARN	T1 APPROACHES TO LEARNING
T1RARSGE	T1 REC GENERAL KNOWLEDGE ARS SCORE
T1RARSLI	T1 REC LITERACY ARS SCORE
T1RARSMA	T1 REC MATH ARS SCORE
W1INCOME	W1 INCOME (IMPUTED)
W1SESL	W1 CONTINUOUS SES MEASURE
WKSESL	WK CONTINUOUS SES MEASURE

Appendix D: Final List of 66 Covariates

Table D1

Final 66 Covariates: Scale Variables

Variable Name	Mean	SD
BIRTH WT	117.17	21.55
C1BMI	16.27	2.21
C1FMOTOR	5.77	2.06
C1HEIGHT	44.65	2.16
C1R4MSCL	25.36	8.66
C1RGSCAL	21.37	7.5
P1CHLAUD	14.38	17.63
P1CHLBOO	70.8	58.97
P1FTHGRD	12.44	4.54
P1HIG_1	14.4	3.38
P1HMAFB	23.27	5.34
P1HMAGE	32.95	6.81
P1HMLTOM	1.36	0.85
P1HTOTAL	4.52	1.37
P1IMPULS	1.99	0.69
P1LEARN	3.1	0.48
P1MTHGRD	12.32	4.08
P1NUMARR	1.23	0.56
P1OVER18	2.02	0.68
P1RDAYPK	3.89	1.49
P1SADLON	1.55	0.41
P2INCOME	44928.63	34679.74
T1EXTERN	1.62	0.64
T1INTERN	1.53	0.53
T1INTERP	2.97	0.63
T1RARSMA	2.53	0.8
W1INCOME	16381.24	6815.96

(Appendix D Continues)

Table D2

Final 66 Covariates: Ordinal Variables

Variable Name	Maximum	Median
P1BEHAVE	4	2
P1BUILD	4	2
P1CHLPIC	4	4
P1CHORES	4	4
P1CHREAD	4	3
P1GAMES	4	3
P1HELPAR	4	3
P1HIGHSC	8	3
P1HSCALE	5	1
P1NATURE	4	2
P1PRIMPK	8	5
P1READBO	4	3
P1RMOPK	4	4
P1SINGSO	4	3
P1SPORT	4	3
P1TELLST	4	3

(Appendix D continues)

Table D3

Final 66 Covariates: Nominal Variables

Variable Name	Percent[a]	Coded "1" Represents
C1SCREEN	.87	Speaks English in home
C1SPHOME	.91	Does not speak Spanish in home
P1ADLTLV	.80	Responding adults live in home
P1CARNOW	.50	Child not currently in any daycare
P1CENTER	.25	Child never was in center-based daycare
P1CHOOSE	.68	School did not affect selection of home
P1CHSESA	.62	Watched Sesame Street
P1DADOCC	.62	Dad in service profession
P1DISABL	.85	Child does not have disability
P1EARINF	.68	Child does not often have ear infections
P1HDAD	.34	Biological Dad does not live in home
P1HDEMP	.87	Dad works full-time
P1HFAMIL	.26	Does not have two parents in home
P1HMEMP	.46	Mother works full-time
P1HMOM	.07	Biological mother does not live in home
P1HMRACE	.39	Non-white mother
P1HSEVER	.80	Child did not attend Head Start
P1MOMOCC	.42	Mother in service profession
P1PREMAT	.82	Child was not two weeks or more premature
P1STPREP	.31	Child did not attend kindergarten preparation
P1WICCHD	.50	Child not qualified for WIC
P1WICMOM	.55	Mother not qualified for WIC
RACE	.43	Non-white child

Note. All nominal variables were re-coded to two levels.

[a]Percent coded as "1"

Appendix E: Reduced Lists of 17 and 7 Covariates

Table E1

Reduced List of 17 Covariates
C1FMOTOR
C1R4MSCL
C1RGSCAL
C1SPHOME
P1CARNOW
P1CHLBOO
P1DISABL
P1HIGHSC
P1HMAGE
P1HMEMP
P1HMLTOM
P1LEARN
P1MOMOCC
P1SINGSO
T1EXTERN
T1INTERN
T1RARSMA

Table E2

Reduced List of 7 Covariates
C1FMOTOR
C1RGSCAL
C1RSMSCL
C1SPHOME
P1HMAGE
T1EXTERN
T1RARSMA

Appendix F: Value-Added Covariate Parameter Estimates, Fixed Approach

Variable	*b*	SE	*t*	*p*
C1RGSCAL	0.17	0.01	13.32	.000
C1R4MSCL	0.86	0.01	82.09	.000
C1FMOTOR	0.44	0.04	12.22	.000
P1LEARN	0.33	0.15	2.22	.027
P1SADLON	-0.11	0.17	-0.67	.505
P1IMPULS	-0.03	0.10	-0.26	.799
T1RARSMA	0.63	0.11	5.91	.000
T1EXTERN	-0.42	0.13	-3.24	.001
T1INTERN	-0.24	0.14	-1.77	.076
C1HEIGHT	0.06	0.03	1.87	.062
C1BMI	-0.03	0.03	-1.00	.320
P1HMAGE	0.04	0.02	2.80	.005
P1HMAFB	-0.03	0.02	-1.69	.091
P1HMLTOM	0.17	0.12	1.46	.143
P1OVER18	0.08	0.12	0.66	.508
P1HTOTAL	-0.05	0.06	-0.74	.457
W1INCOME	0.00	0.00	1.58	.114
P1CHLBOO	0.00	0.00	1.94	.053
P1CHLAUD	-0.01	0.00	-1.28	.202
P1NUMARR	-0.16	0.12	-1.33	.183
P1RDAYPK	0.02	0.05	0.29	.772
P1HIG_1	0.02	0.03	0.56	.574
P1FTHGRD	-0.01	0.02	-0.74	.459
P1MTHGRD	0.01	0.02	0.20	.842
P2INCOME	0.00	0.00	-1.52	.130
BIRTH WT	0.00	0.00	1.00	.317
T1INTERP	0.23	0.14	1.57	.117
P1PRIMPK	0.01	0.03	0.23	.818
P1READBO	0.10	0.09	1.02	.309
P1TELLST	-0.13	0.08	-1.76	.079
P1SINGSO	-0.12	0.07	-1.69	.091
P1HELPAR	-0.03	0.08	-0.35	.724
P1CHORES	0.11	0.08	1.49	.137
P1GAMES	0.13	0.08	1.54	.124
P1NATURE	-0.06	0.08	-0.76	.435
P1BUILD	0.03	0.07	0.37	.710

(Table continues)

(continued)

Variable	*b*	SE	*t*	*p*
P1SPORT	0.04	0.07	0.48	.634
P1CHLPIC	0.07	0.09	0.80	.423
P1CHREAD	-0.10	0.08	-1.25	.211
P1RMOPK	-0.04	0.08	-0.57	.567
P1BEHAVE	-0.09	0.11	-0.82	.413
P1HSCALE	-0.03	0.08	-0.43	.667
P1HIGHSC	-0.10	0.04	-2.22	.027
RACE	-0.33	0.26	-1.24	.214
P1HMEMP	0.33	0.15	2.21	.027
P1HDEMP	0.00	0.19	0.00	.997
P1HMRACE	0.15	0.27	0.54	.593
P1MOMOCC	-0.32	0.14	-2.25	.025
P1DADOCC	-0.08	0.15	-0.51	.609
P1CARNOW	0.28	0.15	1.88	.061
P1CENTER	-0.19	0.20	-0.95	.343
P1DISABL	0.34	0.18	1.82	.068
P1HMOM	-0.45	0.30	-1.49	.136
P1HDAD	0.00	0.20	-0.01	.988
P1HFAMIL	-0.02	0.23	-0.07	.949
C1SCREEN	-0.59	0.36	-1.61	.107
C1SPHOME	1.18	0.40	2.98	.003
P1STPREP	-0.12	0.14	-0.85	.397
P1CHOOSE	-0.11	0.14	-0.78	.437
P1CHSESA	-0.18	0.14	-1.30	.194
P1HSEVER	0.10	0.19	0.55	.585
P1PREMAT	-0.24	0.18	-1.33	.183
P1EARINF	0.03	0.14	0.23	.822
P1ADLTLV	0.00	0.16	-0.02	.987
P1WICMOM	-0.16	0.22	-0.72	.474
P1WICCHD	0.16	0.23	0.71	.480

Appendix G: Variables with Statistically Significant Imbalance after Matching

Table G1

66 Matching Variables, 435 Teachers, Number of Class-Control Pairs with Mean or Proportion Differences Statistically Significant at .05

Variable	Before Matching	1:1 Match	1:2 Match	1:5 Match	1:20 Match
C1RGSCAL	147	3	1	4	6
C1R4MSCL	100	3	1	1	6
C1FMOTOR	81	0	0	1	4
P1LEARN	38	3	1	1	1
P1SADLON	54	0	0	1	2
P1IMPULS	60	1	1	0	1
T1RARSMA	177	2	3	7	21
T1EXTERN	97	3	2	5	7
T1INTERN	136	3	3	2	15
C1HEIGHT	46	2	1	0	3
C1BMI	49	3	2	0	1
P1HMAGE	75	1	1	3	5
P1HMAFB	118	1	1	2	4
P1HMLTOM	256	1	3	8	58
P1OVER18	61	1	1	0	8
P1HTOTAL	45	3	2	1	3
W1INCOME	78	3	1	0	1
P1CHLBOO	107	4	1	4	12
P1CHLAUD	85	0	2	2	8
P1NUMARR	24	0	0	0	0
P1RDAYPK	78	3	0	0	1
P1HIG_1	131	0	0	2	3
P1FTHGRD	99	1	1	1	3
P1MTHGRD	84	0	0	1	3
P2INCOME	162	1	1	2	12
BIRTH WT	32	2	0	0	0
T1INTERP	148	4	2	4	9
P1PRIMPK	70	2	0	0	4
P1READBO	68	2	0	1	3
P1TELLST	32	2	1	1	2
P1SINGSO	43	0	0	2	5
P1HELPAR	40	2	0	0	1
P1CHORES	47	1	0	1	6

(Table continues)

Table G1(continued)

66 Matching Variables, 435 Teachers, Number of Class-Control Pairs with Mean or Proportion Differences Statistically Significant at .05

Variable	Before Matching	1:1 Match	1:2 Match	1:5 Match	1:20 Match
P1GAMES	25	1	1	1	2
P1BUILD	35	1	0	0	0
P1SPORT	44	2	0	0	3
P1CHLPIC	54	1	0	2	7
P1CHREAD	46	1	0	2	4
P1RMOPK	49	1	0	3	7
P1BEHAVE	43	0	0	0	2
P1HSCALE	72	2	0	2	3
P1HIGHSC	45	1	0	0	2
RACE	177	0	0	0	1
P1HMEMP	20	0	0	0	0
P1HDEMP	16	0	0	0	0
P1HMRACE	156	0	0	0	0
P1MOMOCC	30	1	0	0	0
P1DADOCC	63	0	0	0	0
P1CARNOW	23	0	0	0	0
P1CENTER	29	0	0	0	0
P1DISABL	13	0	0	0	0
P1HMOM	15	0	0	0	1
P1HDAD	47	1	0	0	0
P1HFAMIL	40	0	0	0	1
C1SCREEN	48	0	0	0	3
C1SPHOME	25	0	0	0	1
P1STPREP	34	0	0	1	1
P1CHOOSE	35	0	0	0	0
P1CHSESA	33	0	0	0	0
P1HSEVER	49	0	0	0	2
P1PREMAT	8	0	0	0	0
P1EARINF	7	0	0	0	0
P1ADLTLV	16	0	0	0	0
P1WICMOM	106	0	0	0	1
P1WICCHD	114	0	0	1	1

Table G2

17 Matching Variables, 435 Teachers, Number of Class-Control Pairs with Mean or Proportion Differences Statistically Significant at .05

Variable	Before Matching	1:1 Match	1:2 Match	1:5 Match	1:20 Match
C1RGSCAL	147	4	0	0	0
C1R4MSCL	100	1	1	0	0
C1FMOTOR	81	5	0	0	0
T1RARSMA	177	10	6	0	3
T1EXTERN	97	4	0	0	0
C1SPHOME	25	0	0	0	0
P1HMAGE	75	4	0	0	0
P1MOMOCC	30	0	0	0	0
P1LEARN	38	1	0	0	0
P1HIGHSC	45	4	0	0	0
P1HMEMP	20	0	0	0	0
P1CHLBOO	107	5	0	1	0
P1CARNOW	23	0	0	0	0
P1DISABL	13	0	0	0	0
T1INTERN	136	4	0	0	2
P1SINGSO	43	3	2	0	0
P1HMLTOM	256	1	0	1	11