TEACHER EFFECTIVENESS AND CAUSAL INFERENCE IN OBSERVATIONAL STUDIES

Roderick A. Rose

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Public Policy

Chapel Hill
May 2013

Approved by:

Christine P. Durrance

Gary T. Henry

John C. Scott

Judith L. Meece

Michael E. Woolley

ABSTRACT

RODERICK A. ROSE: Teacher Effectiveness and Causal Inference in Observational Studies
(Under the direction of Christine P. Durrance, PhD)

An important target of education policy is to improve overall teacher effectiveness using evidence-based policies. Randomized control trials (RCTs), which randomly assign study participants or groups of participants to treatment and control conditions, are not always practical or possible and observational studies using rigorous quasi-experimental or comparison group methods must frequently be used. Each of the three studies in this dissertation studies a facet of teacher effectiveness that can be used to inform policy while also contributing to scholarship in the area of causal inference from observational data. First, estimating or measuring individual-level teacher effectiveness using student value-added requires models that are robust to observational data. Second, teacher participation in a teaching intervention does not necessarily imply compliance with the treatment. The factors that interfere in compliance are non-random such that understanding the role of the hypothesized process in an RCT requires observational methods. Third, if observational methods such as propensity score analysis offer credible evidence of causality, they may be useful for examining efficacious interventions that are taken to scale in non-random settings. These issues place observational designs and methods at the center of efforts to study evidence-based teaching and teaching practices.

DEDICATION

This dissertation is dedicated to the memory of Dr. Gary L. Shaffer, a devoted school social worker, a tireless advocate for the rights and well-being of North Carolina's children, and a friend.

Gary was an extraordinary person. While educating and licensing school social workers, fighting against corporal punishment in schools, advocating for anti-bullying legislation, and fighting the cancer that ultimately took him from us four years ago, he found the time to enjoy a few beers with his friends and colleagues at the School of Social Work.

On one memorable occasion, Steve Day, Gary and I went to the Carolina Brewery during the holiday season. The brewery was serving a beer called "Santa's Secret." We all asked to sample this beer. It wasn't their best work. After tasting, one of us (Steve or I) asked what the "secret" in Santa's Secret was. Gary responded, "Santa's secret is, don't drink this beer."

I am reminded of this every time I think of Gary. But more importantly the devotion to the children of North Carolina shown by people like Gary is one reason why I chose to focus on education policy and I dedicate this work in his memory as a tribute to his legacy.

ACKNOWLEDGEMENTS

This dissertation is my most significant personal and professional accomplishment, and its genesis dates all the way back to 1997, when I applied to UNC to get a doctorate in *economics*. Over the 16 years between then and now I have received help, assistance, and encouragement from a lot of really great people.

The first person I want to thank is the chair of my advisory committee, Christine Durrance. Christine, an economist and assistant professor in the department, stepped up to help me in a big way when I asked her to take over as chairperson in December 2012 in the eleventh hour of this project. Christine, despite not being an education policy researcher, did a remarkable job of helping me to finalize this dissertation. She listened, she gave tough but constructive criticism, and she clearly put aside whatever personal and professional feelings she may have had for some of the un-economic things I did in these pages, to see this project through to completion. I am forever indebted to her for her timely help. (Thanks to her, I also remembered that economists are people too, something I admit I had forgotten thanks to my two years in the economics program.)

I am also deeply indebted to my advisor, Gary Henry. Gary first made contact with me in spring 2008 to extend the offer to join the department as a graduate student. Despite steadfastly avoiding him during my first year as a student, he somehow managed to appoint himself my advisor during my second year. After overseeing my program plan and field paper, he provided very capable guidance on this project during the fall of 2012, even after leaving UNC to take a position at Vanderbilt. I appreciate his guidance, his assistance, his

genuinely collegial nature, and most of all, his patience. I had some of my crazier moments while his advisee.

I am also grateful to the assistance of Judith L. Meece and John C. Scott, members of my advisory committee, for giving their valuable time and expertise to this project. I originally asked John to be on my committee because I was going to propose a qualitative study for one chapter. Ultimately, I decided against the qualitative study, but stuck with John. I first met Judith when working on the Race to the Top program in January 2011, and knowing of her work in motivation research I mentioned to her *that day* that I wanted her to consider being on my committee. When she accepted I don't think she quite knew what she was getting into. Judith persevered and provided excellent feedback on this project and I am hopeful we will continue to work together as I value her expertise, friendship and guidance.

There are many people who provided "indirect support" to this dissertation and my work on the doctorate over the past few years. I am grateful to have had the world's best cohort, who convinced me that nobody should go through graduate school alone. To Kevin Bastian, Jade Marcus-Jenkins, Shanyce Campbell, Jae Hee Park, and our den mother, Shelley Golden, thanks for persisting, and thanks for making my memories of our shared struggles together in The Program fond ones.

Supportive faculty in the Department of Public Policy provided additional informal guidance on these papers, including Sudhanshu Handa, Jeremy Moulton, and Doug Lauen. I would also like to thank supportive faculty and friends in the UNC School of Social Work and social work research communities for the opportunities that they gave me over the years, including Gary L. Bowen, Dennis K. Orthner, Susan L. Parish, Mark W. Fraser, Shenyang Guo, and Dean F. Duncan, who was the first researcher I worked with at the school when I

started in 1999, and who I currently work with. It was really under their guidance that I learned how to do quality research and how to write and get my work published. Gary, Mark, Dennis and Susan wrote letters for my application to the department, and Mark played a pivotal role in encouraging me to pursue a doctorate.

A large part of this dissertation was made possible by my involvement as Evaluator on the CareerStart evaluation project from 2008 to the present. I am grateful to my colleagues on this project including Patrick T. Akos, Hinckley Jones-Sanpei, Micaela Mercado, Julie McCann and Jay Unick. Another large part of this dissertation was made possible by a project undertaken at the Carolina Institute for Public Policy, and I am grateful to former colleagues there and on the Race to the Top project for their support, including Charles Thompson, Trip Stallings, and Julie Marks.

Several people provided specialized support to enable me to complete this project. I appreciate the assistance of Terri Snyder, editor extraordinaire, who spent many hours to make sure that this document conformed to APA formatting. Mary Leigh Creedon, Student Services Manager in the department, capably answered my frequent questions about the dissertation process, and also helped prepare the room for my final defense. Elizabeth D'Amico, Business Manager at the Carolina Institute for Public Policy, supported efforts that I made to have this project funded.

Finally, I would like to thank the friends and family who have supported my scientific career aspirations. First, I want to thank Michael E. Woolley, friend, colleague, mentor, musical collaborator, CareerStart Principal Investigator, and advisory committee member for his supportiveness and guidance the last several years. I have enjoyed working with Mike and I hope our continued professional collaborations can be as fruitful as they are enjoyable.

I also thank Ken Kastleman, who always listened to me and dispensed helpful advice at critical times during my years as a doctoral student; my fellow beer nerd and Social Work colleague Steve Day; my long-time friend Ross Stephens, who has inspired me to reach higher ever since our days together on our high school cross-country and track teams; other graduate students in the Department of Public Policy and School of Education that I've had the pleasure to know or work with over the years; and the numerous supportive people I have met and come to know in the PM:X and ProgPowerUSA communities.

I would like to also thank my parents Roderick S. and Judith Rose, Lisa's mother Laraine McCaskill, and my brother Kyle Rose for their supportiveness the past several years. I recall in particular the many times between 2000 and 2007 that my father asked me if I would be going back to get that elusive PhD.

Finally, I want to thank my family, Jessica Loehman and Lisa McCaskill, for everything they've done and all the support they've given me over the years. I love them both and am very happy that they are both pursuing their own passions just as I have. In a couple of years, Jessica will be standing where I stand now, with a PhD and a bright future. Lisa, my partner of almost 10 years, lived the graduate student experience with me these last five years, picking up the slack at home when I could not keep up, lending an ear when I needed to complain or worry, and generally keeping me grounded and positive throughout the entire unforgiving but memorable experience.

Roderick A. Rose
March 29, 2013

TABLE OF CONTENTS

LIST OF TABLES xiii

LIST OF FIGURES xvii

INTRODUCTION 1

Background and Significance 2

Rubin Causal Model 7

CHAPTER

1. THE CONTEMPORANEOUS INFLUENCE OF CLASSROOM PEER ASSIGNMENT AND STUDENT-TEACHER INTERACTIONS ON TEACHER VALUE-ADDED MODEL ESTIMATION 12

The Role of Quality Teaching in Student Achievement 15

Teacher Evaluation using Value-Added Models 18

Advantages over Existing Approaches 21

Concerns About VAMs as Evaluation Tools 24

Measurement Validity and Reliability 26

Attributing Student Test Score Changes to Teachers 28

The Rubin Causal Model: Assumptions and Violations 29

Stable Unit Treatment Value Assumption 31

Unconfoundedness 31

The Classroom Effect in the Context of the RCM 33

Aggregate Pre-Assignment Characteristics 34

Student Peer Interactions 36

Teacher-Student Interactions 38
The Classroom Effect as a TVA Disturbance 40
Proposed Solutions to the Classroom Effect Disturbance 42
Objective and Research Questions 44
Methods 45
Value-Added Models 45
Actual North Carolina Data 49
Data Generation Process 50
Methods to Minimize Classroom Effects Disturbances 56
Comparison Methods and Criteria 58
Results 60
Variance Decomposition & SUTVA 60
Correlated Covariates and Unconfoundedness 62
Proposed Solutions 63
Discussion 66
Limitations 68
Implications for Teacher Evaluation Policy 70
Implications for Further Research 79
Concluding Comments 83
2. CAREER RELEVANT INSTRUCTION AS A CAUSAL PROCESS IN THE CAREERSTART TEACHING INTERVENTION PROGRAM 107
Engagement and Motivation 110

School Engagement as a Learning Process 111
Motivational Teaching Practices 112
Relevance and Motivation 113
Motivational Theories 114
Multiple Motivational Pathways for Relevance 116
Relevance as an Intervention in Instructional Practice 118
The Challenges of Decontextualized Instruction 119
Evidence for Relevance Oriented Teaching Strategies 122
CareerStart: A Career Relevant Instruction Intervention 125
Evidence for CareerStart and CRI 126
Method 130
Design 130
Data Sources 134
Sample 134
Measures 135
Missing Data 140
Analysis 142
Robustness Analyses 146
Results 147
CRI Measurement Results 147
Analysis Results 149
Discussion 153

Limitations 158

Implications for Further Research 160

Implications for Education Policy 164

Conclusion 168

3. A NON-RANDOMIZED EFFICACY TRIAL REPLICATION OF A TEACHING PRACTICE INTERVENTION USING PROPENSITY SCORE WEIGHTING 214

CareerStart: Engagement, Motivation and Relevance 217

CareerStart 218

Evidence of the Efficacy of CareerStart 222

Intervention Scalability 224

Scaling from Efficacy to Effectiveness 225

Extant Research on the Scalability of Motivational Instructional Practices 226

Taking CareerStart to Scale: An Effectiveness Trial 228

Non-Random Replication of CareerStart 231

Propensity Score Analysis and the Rubin Causal Model 232

Propensity Score Modeling 235

Propensity Score Methods and Unconfoundedness 237

PSA as a Replication Method 240

Study Objectives and Research Questions 242

Methods 243

PSA Design 243

Data and Sample 245

Measures 247
Data Analysis 250
Results 258
Propensity Score Model 258
Balance 259
Analysis Model 260
Analysis Model Robustness 264
Discussion 264
Limitations 267
Implications for Policy 270
Implications for Research 273
Conclusion 278
APPENDIX 1. ESTIMATION OF COMPOSITE IN URM 349
APPENDIX 2. CAREERSTART EIGHTH GRADE MATHEMATICS LESSON 350
REFERENCES 352

LIST OF TABLES

Table 1.1. Summary of Model 86

Table 1.2. List of Variables in Actual NC Data used in Calibrations and HLM4 Analysis 87

Table 1.3. Calibration of Variance Component Inputs 89

Table 1.4. Mean Squared Error, Variance Decomposition Models 90

Table 1.5. Correlation of Estimated Teacher Effect with True Effect at Varying Levels of Classroom Variance 91

Table 1.6. Percent Agreement between TVA and True Teacher Effect on Identifying Bottom 5% of Teachers 92

Table 1.7. Number and Proportion of Teachers Falsely Identified as Ineffective 93

Table 1.8. True Scores of Teachers Falsely Identified as Ineffective 94

Table 1.9. Mean Squared Error 95

Table 1.10. Spearman Rank Order Correlation by Correlation between Teacher and Classroom 96

Table 1.11. Percent Agreement between TVA and True Estimate on Bottom 5% of Teachers 97

Table 1.12. Proportion and Number of Teachers Falsely Identified as Ineffective Assuming 9,000 Fifth Grade Teachers 98

Table 1.13. True Z Scores of Teachers Falsely Identified as Ineffective 99

Table 1.14. Correlation of Estimated Teacher Effect with True Effect at Varying Levels of Classroom Variance 100

Table 1.15. Percent Agreement between TVA and True Teacher Effect on Identifying Bottom 5% of Teachers 101

Table 1.16. Number and Proportion of Teachers Falsely Identified as Ineffective by Classroom Proportion of Variance 102

Table 1.17. True Scores of Teachers Falsely Identified as Ineffective 103

Table 1.18. Correlated Covariate Simulation: 2 vs. 4 Cohorts 104

Table 1.19. HLM4 Model Results: Variance Decomposition Models 105

Table 1.20. True Scores, Proportion, and Number of Teachers
Falsely Identified as Ineffective (HLM4 and CCOV Models) 106

Table 2.1. Sample Characteristics (N = 2866).....175

Table 2.2. CRI Binary Variable 179

Table 2.3. Decomposition of Variance in CRI 180

Table 2.4. CRI Measurement Model for Language Arts: Factors
Explaining Variation in Student Perceptions of 8th Grade CRI 181

Table 2.5. CRI Measurement Model for Math: Factors Explaining
Variation in Student Perceptions of 8th Grade CRI.....183

Table 2.6. CRI Measurement Model for Science: Factors Explaining
Variation in Student Perceptions of 8th Grade CRI.....184

Table 2.7. CRI Measurement Model for Social Studies: Factors
Explaining Variation in Student Perceptions of 8th Grade CRI 186

Table 2.8. IVE Models without Teacher-Level Imputation,
Different Covariate Specifications.....188

Table 2.9. IVE Models with Teacher-Level Imputation,
Different Covariate Specifications.....189

Table 2.10. IVE Main Models with Observations Having Missing
Teacher Covariates Deleted 190

Table 2.11. No IVE: Three-Level Rich Covariate Models without
Teacher Level Imputation, Different Covariate Specifications 191

Table 2.12. IVE Models without Teacher Level Imputation,
Different Covariate Specifications.....192

Table 2.13. EOG Math 8th Grade IVE/LATE 193

Table 2.14. EOG Reading 8th Grade IVE/LATE 195

Table 2.15. Engagement, Math CRI IVE/LATE 197

Table 2.16. Engagement, LA CRI IVE/LATE 199

Table 2.17. Engagement, Science CRI IVE/LATE 201

Table 2.18. Engagement, SS CRI IVE/LATE 203

Table 2.19. Valuing, Math CRI IVE/LATE 205

Table 2.20. Valuing, LA CRI IVE/LATE.....207

Table 2.21. Valuing, Science CRI IVE/LATE 209

Table 2.22. Valuing, SS CRI IVE/LATE 211

Table 2.23. Models for Academic Press and Preparation for Future Schooling as Endogenous Variable (Replaces CRI).....213

Table 3.1A. Cohort Entry and Years Observed (Aligned by Treatment Year) 282

Table 3.1B. Cohort Entry and Grade Levels Observed (Aligned by Treatment Year) 282

Table 3.2. Heterogeneity between Counties 283

Table 3.3. Variables used in Estimation 286

Table 3.4. Univariate Statistics 293

Table 3.5. Model Variations 298

Table 3.6. Correlations between Outcome Proxies and Candidates for PSM Covariates 299

Table 3.7. Propensity Score Model—Imputed Data (M = 5) 302

Table 3.8. Balance before PSA Model (Variables in PSA Model).....305

Table 3.9. Balance Before PSA Model (Variables Not in PSA Model—Supplement).....308

Table 3.10. Balance After PSA Model (Variables in PSA Model) 310

Table 3.11. Balance After PSA Model (Variables Not in PSA Model—Supplement).....313

Table 3.12. Sample Description after Weighting.....315

Table 3.13A. Treatment Effects (Questions) 320

Table 3.13b. Credibility of PSA Design 321

Table 3.14. Math (Overall Treatment Effect—Imputed Data) 323

Table 3.15. Math (Difference in PSA Treatment Effect between
RCT Control and Non-RCT—Imputed Data).....327

Table 3.16. Math (Treatment Effects for RCT Control,
Non-RCT and RCT Treatment—Imputed Data) 331

Table 3.17. Reading (Overall Treatment Effect—Imputed Data) 335

Table 3.18. Reading (Difference in PSA Treatment Effect
between RCT Control and Non-RCT—Imputed Data) 339

Table 3.19. Reading (Treatment Effects for RCT Control,
Non-RCT and RCT Treatment—Imputed Data) 343

Table 3.20. Intensity of Implementation in RCT and
PSM Evaluation Periods: Number of CareerStart Lessons Used By
School and Subject.....347

LIST OF FIGURES

Figure 1.1. Shared variance between student growth and classroom and teacher effects. 85

Figure 1.2. Mean squared error, variance decomposition model.....90

Figure 1.3. Spearman correlation by classroom variance. 91

Figure 1.4. Percent agreement on 5th percentile by classroom variance. 92

Figure 1.5. True scores of teachers falsely identified as ineffective. 94

Figure 1.6. Mean squared error, correlated covariate models.....95

Figure 1.7. Spearman rank order correlation by level of correlation between classroom and teacher. 96

Figure 1.8. Percent agreement, bottom 5% of teachers. 97

Figure 1.9. True Z scores of teachers above cutoff with TVAEs below cutoff by level of correlation between classroom and teacher. 99

Figure 1.10. Spearman correlation by classroom variance. 100

Figure 1.11. Percent agreement on 5th percentile by classroom variance. 101

Figure 1.12. True scores of teachers falsely identified as ineffective. 103

Figure 2.1. CareerStart theory of change model. 170

Figure 2.2. Kernel density engagement. 171

Figure 2.3. Kernel density valuing. 171

Figure 2.4. Kernel density math. 172

Figure 2.5. Kernel density reading. 172

Figure 2.6. Kernel density LA CRI. 173

Figure 2.7. Kernel density math CRI. 173

Figure 2.8. Kernel density science CRI. 174

Figure 2.9. Kernel density social studies CRI. 174

Figure 3.1. Box plot, raw propensity score. 279

Figure 3.2. Box plot, propensity scores trimmed
at .05 (comparison) and .95 (tx).....279

Figure 3.3. Kernel density function, raw propensities,
all participants.....280

Figure 3.4. Kernel density function, trimmed propentities,
all participants.....280

Figure 3.5. Kernel density function, raw propensities, by
treatment condition. 281

Figure 3.6. Kernel density function, trimmed propensities,
by treatment condition. 281

INTRODUCTION

Teacher effects are widely seen by education policy scholars as the most important school input to student learning. If this is true, then an important target of education policy must be to improve overall teacher effectiveness and ensure an equitable distribution of effective teaching, outcomes that can be manipulated through targeted evidence-based policy interventions. These policies are labeled evidence-based because rigorous scientific methods are used to rule out plausible alternative explanations for observed effects, allowing researchers to ascribe causality to the policy. Randomized control trials (RCTs), which randomly assign study participants or groups of participants to treatment and control conditions, are the most rigorous design for allowing causal inferences and contributing to evidence-based policy. However, RCTs are not always practical or possible and observational studies using rigorous quasi-experimental or comparison group methods must be used. These methods include fixed effects (Allison, 2009), instrumental variable estimation (Angrist, Imbens & Rubin, 1996); propensity scores matching and weighting (Rosenbaum & Rubin, 1983); regression discontinuity (Imbens & Lemieux, 2009); and interrupted time series (Bloom, 2003). The relative merit of observational studies is fairly controversial and remains the subject of scholarly work in education and other policy fields (see Cook & Steiner, 2009; Pirog, 2009).

Each of the three studies in this dissertation studies a facet of teacher effectiveness that can be used to inform policy while also contributing to scholarship in the area of causal inference from observational data where randomization is not practical or possible. First,

estimating or measuring individual-level teacher effectiveness, typically obtained using models that aggregate student achievement data and attribute these aggregated effects to their teachers, requires models that are robust to observational data because teachers are generally not assigned randomly to students. Second, if observational methods are used in teaching practice intervention studies, as they typically are when there are difficulties in obtaining sufficiently large and heterogeneous study samples, they must offer credible evidence of causality. Demonstrating the equivalence of experimental and quasi-experimental studies in a population subjected to a particular treatment can help this effort. Third, the underlying mechanism promoting effectiveness, typically seen as important to developing a base of knowledge on effective teaching practices, requires observational methods, even in settings where a teaching practice intervention is randomized to groups of teachers such as in an RCT. Participating in the intervention does not necessarily imply compliance with the treatment and the factors that interfere in compliance are non-random. These issues place observational designs and methods at the center of efforts to study evidence-based teaching and teaching practices.

Background and Significance

The need for rigorous scientific methods to establish evidence-based policy is usually inferred to mean randomized experiments. The objective of a social or policy experiment is to determine the extent to which a manipulable condition such as a policy *causes* an intended effect, on average, in the target population (Heckman & Smith, 1995). This requires the researcher to not only show that an effect has been produced during the study period but that other plausible explanations for the effect have been ruled out, which can be onerous (Shadish, Cook & Campbell, 2002). An experimental study in which study participants have

been randomized to treatment and control conditions simplifies the latter condition because participants in the two conditions have a high probability of being similar to each other (Shadish, Cook & Campbell, 2002). Randomly assigning members of the sample to each condition means that in practice no pretreatment characteristic is taken into account in making assignment. In any given experiment, this process can produce chance differences between treatment conditions, but over repeated versions of this experiment no characteristic would be expected to emerge as significantly imbalanced between the two groups. Because of this expectation the treatment effect can be viewed as a simple difference between the average effect of those in the treated and non-treated groups as follows (Morgan & Winship, 2007):

$E[d] = E[Y_1] - E[Y_0] = E[Y_1 - Y_0]$ with 1 = treated, 0 = control.

Randomization is widely seen both in the scientific and policy research communities as the "gold standard" of research and evaluation. Federal mandates such as No Child Left Behind require that for policy to be considered evidence-based randomized experiments are preferred (Shadish & Cook, 2009). Education and education policy researchers responded by increasing their use of randomized experiments (Constas, 2007). The simplicity by which randomization works to promote causality makes it transparent and thus appealing to the researcher and policymaker alike, leading to political support for policies based on experimental evidence (Balducci & Wandner, 2009; Heckman & Smith, 1995). When studying teacher effectiveness, for example, a random assignment of teachers to students would make us more confident that typical student learning advantages such as parental resources are randomly distributed among teachers, and as a consequence the average end-of-grade test score for the students taught by each teacher could be interpreted as an indicator of

that teacher's effectiveness (Nye, Konstantopoulos, & Hedges, 2004). When studying the effects of a teaching practice treatment, random assignment of teachers to conditions – a condition in which they receive the treatment and a condition in which they do not – would increase our confidence that the observed difference in average student performance in each condition could be interpreted as the effect of the treatment.

However, the conditions for these types of studies cannot always be met, making randomization impractical if not impossible. First, in many settings in which education policy researchers conduct their research, there are multiple social or political barriers to conducting randomized experiments (Cook, 1999). As I earlier noted, students and teachers are not typically randomly assigned to each other (Nye, Konstantopoulos, & Hedges, 2004). The actual assignment mechanisms used may be tailored to meet the needs of students or the wishes of their families, such as compensatory assignments in which lower-performing students are assigned to more effective teachers, or tracking systems in which more effective students and teachers are matched to each other. The evidence suggests that observed student achievement and teacher effectiveness are positively associated (Clotfelter, Ladd & Vigdor, 2006). Although researchers have in the past succeeded in convincing schools to randomize the assignment mechanism for a short time (the Tennessee STAR study being the most well-known example; Nye, Hedges, & Konstantopoulos, 1999), in most settings this is not the case. Consequently, to estimate teacher effects from aggregate student data and attribute those effects to the teacher, which implies a causal relationship between the teacher's effectiveness and the student outcome, methods that are appropriate for causality in observational studies must be used.

Second, although randomized experiments of teaching practice interventions promote causal inference by averaging over pretreatment characteristics of participants in each treatment group, they do not guarantee generalizability of study findings over all target populations, particularly when conducted on small homogeneous populations (Shadish, Cook & Campbell, 2002). Finally, many in-school intervention studies require that entire classrooms or schools rather than their constituent members be assigned to treatment conditions. These studies, known as a cluster randomized trials, increase the burden on the researcher to satisfy the minimum effective sample size for a given level of statistical power, and ultimately leads to a higher cost of conducting in-school research and a higher risk that a study will be under-powered (Hedges & Hedberg, n.d.; Schochet, 2005). These three conditions often co-exist in randomized experiments conducted in schools due to the challenges faced by researchers in having schools or their gatekeepers at the district level agree to participate in a study design that calls for post-recruitment randomization into treatment conditions, resulting in fewer and more selective volunteer districts. Non-random study samples of more dispersed volunteer participants may present opportunities to address the problems of homogeneity and power.

Finally, it is often the case that direct manipulation of the construct of interest through randomization is not possible. For example, although a packaged treatment can be randomly assigned to schools or teachers enabling a straightforward causal inference of the effects of this treatment, the processes that the treatment as designed is theorized to promote cannot themselves be randomized and may be subjected to multiple nonrandom factors at the district, school, teacher, and classroom levels (Schochet & Chiang, 2009). In the case of a typical teaching practice intervention, the objective of the intervention is to change teachers'

practices to those that are consistent with theory or evidence about practices that better facilitate student learning. Imperfect compliance with the recommendations of the treatment implies that non-random factors effectively mediate the relationship between the treatment and compliance. This manifests as teachers' varying usage of the instruments of the treatment and as variability in the proximal teacher outcomes the treatment is intended to promote (e.g., Rose et al., 2012). Although the treatment itself can be randomized, and questions about the effect of the treatment (often labeled "intent to treat") inferred as causal, these mediating processes (often labeled "treatment on treated") cannot be randomized and must be treated as observational.

In all three of these scenarios a robust causal framework helps to clarify the specific challenges that education policy researchers face when randomization is not possible or practical, including in settings where teaching effectiveness or teaching practice is being studied. A number of causal frameworks are available. This includes the Campbell causal model which has enjoyed a long history in social research (Shadish, Cook & Campbell, 2002), and which is based on an enumeration of the challenges to the internal validity of the causal inference of parameter estimates (Shadish, 2010). The second is the Rubin Causal model, which is a causal framework based on a rigorous statistical model and several key assumptions (Holland, 1986). The third is the Pearl causal model (Pearl, 2000), which is approach largely based on directed acyclic graphs. To be consistent with most of the contemporary policy analysis literature, I rely on the Rubin causal model framework, assumptions and terminology.

Rubin Causal Model

In the Rubin causal model (RCM), also known as the potential outcomes model, the definition of the causal estimand for a treatment (a teacher or a teaching practice intervention) depends on the conditions that are experienced in the absence of the specified cause, known as the counterfactual. Formally, assume that the outcome for student i (with I = 1,…N) under treatment condition j is Y_{ij}, with J possible treatments. In many research settings, such as a typical intervention evaluation, this is dichotomous, with j = 1 for the treated and j = 0 for the control, while in others, such as when examining teacher effectiveness, this may be a many-valued treatment with the potential outcomes represented by a matrix of N students by J teachers (Reardon & Raudenbush, 2009; Morgan & Winship, 2007). Second, the student has one potential outcome Y_{ij} under each possible treatment. At most one of these potential outcomes can be realized, as each student can only be assigned to one condition (the fundamental problem of causal inference; Holland, 1986). Consequently, the treatment effect is defined as a function of the distributions of students assigned to treatment j and the students under the other condition or conditions, usually the average treatment effect for treatment j (ATE_j), comprised of the difference between students observed under assignment to treatment j and those assigned to the other conditions.

The assumptions of the RCM highlight the primary benefit from randomization, and help to illustrate that although randomization is seen as the gold standard, it is certainly not perfect. In its most general form, the RCM requires the following assumptions for the ATE_j to be inferred as a causal effect: (1) each student has a potential outcome under each treatment condition in the population (manipulability); (2) the potential outcome under each condition is independent of the assignment of other participants (the stable unit treatment

value or SUTVA assumption); and (3) unconfounded (also known as ignorable or exogenous) assignment to treatment conditions, which implies that participants' assignment to treatment is not associated with the expectation of their outcome under treatment (Morgan & Winship, 2007). The RCM clarifies that the process of randomization, which provides that knowing a participant's treatment assignment offers no information about that participant's potential outcome in any condition, primarily addresses the unconfoundedness assumption (Shadish, Cook & Campbell, 2002). Randomization also makes manipulability a plausible assumption, as the assignment of participants to each treatment condition is under the control of the researcher. However, randomization does not address SUTVA. Yet SUTVA can be violated in any setting in which participants interact with each other under each treatment condition, making it difficult to ascertain what part of an observed effect is due to the treatment and which part may be due to the grouping-dependent interaction. SUTVA, although plausibly violated in many randomized experiments, must simply be assumed.

In observational settings, all three of these assumptions come into question. Manipulability may not be observed in practice. In the study of teacher effectiveness, for example, it may be difficult to support the assumption that students can be assigned to any teacher in any school, as students from disadvantaged backgrounds tend to be educated in poorer schools. Observational studies, like RCTs, make no claims to the plausibility of SUTVA. Finally, the RCM clarifies that the plausibility of a causal estimate from an observational study depends on the extent to which proper observational methods approximate the unconfounded treatment assignment of a randomized design in the population. In fact, methods that rely strictly on a rich set of covariates may be sufficient provided they meet the requirement for unconfoundedness.

Proposed Studies

In this dissertation, I undertook three studies that contribute to the education policy research literature. These studies (1) help to reveal an important confounding issue in the estimation of teacher effectiveness, (2) help to validate the logic model and reveal the processes operating in an evidence-based teaching practice intervention, and (3) replicated the findings of an RCT of this teaching practice intervention using observational methods. Each study also makes a unique contribution to the education policy literature regarding causal inference from observational study designs.

Chapter 1. The first chapter examines questions related to the estimation of value added models (VAMs) for estimating teacher effectiveness for evaluating teachers, in the presence of confounding effects from non-random assignment of students and teachers to classrooms and from student-student and student-teacher interactions. These factors violate the SUTVA and unconfoundedness assumptions of the RCM (Rubin, Stuart & Zanutto, 2004). This study uses simulated data to understand the extent to which the classroom and peer effects bias teacher estimates when no explicit controls for these effects are included and whether certain VAMs are better than others at minimizing this bias, and to determine if there are methods that can be used to adjust for the peer and classroom effects. This study contributes to the body of work on teacher effect estimation for research and teacher evaluation systems, demonstrating the challenges that must be overcome to use VAMs in teacher evaluation systems, and suggests approaches that make best use of the limited and imperfect information that VAMs provide (Henry, Rose & Lauen, under review; Amrein-Beardsley, 2008; Gordon et al., 2006).

Chapter 2. The second chapter examines the logic model of a teaching practice intervention, CareerStart, to determine whether career relevant instruction (CRI), the use of career examples by teachers to associate lesson content and school tasks with something of value to students – their futures, and in particular, the types of careers they might be likely to have – promotes higher student engagement and achievement. Demonstrating such an effect would provide strong evidence for the final leg of the logic model that supports that CareerStart promotes higher use of career examples by teachers (in math; Rose et al., 2012), and higher student valuing of school (emotional engagement; Orthner et al., in press) and math achievement (Woolley et al., under review). In this study I use non-random methods to show that CRI promotes student valuing of school. Because CRI could not itself be randomly assigned, because it is subjected to non-compliance due to non-random factors at the district, school or principal, teacher, and classroom levels, the observed estimate of a unit of CRI cannot generally be interpreted as a causal estimate. Consequently, an instrumental variable framework known as local average treatment effects (Angrist, Imbens & Rubin, 1996) in which the assignment to CareerStart was used as the instrument for CRI was used. This study not only contributes to the investigation of the processes by which CareerStart affects student achievement and engagement, but also contributes to the larger body of work on local average treatment effects as a means of estimating the processes operating in teaching practice interventions (Gennetian, Magnuson & Morris, 2008).

Chapter 3. The third chapter presents a quasi-experimental replication study of a randomized control trial (RCT) of CareerStart, which was shown to improve student achievement on end-of-grade math tests but not end-of-grade reading (Woolley et al., under review). The replication study is based on a non-random assignment of CareerStart to 20

schools in North Carolina. I use this opportunity to test the scalability of the program to populations and contexts outside of the RCT. I use propensity score analysis (PSA) to assign weights to these 20 schools and all other middle schools in North Carolina such that treatment assignment is rendered conditionally random. The challenge of propensity score studies is modeling the assignment process to produce effectively unconfounded assignment to treatment, and the combination of matching and regression methods using rich covariates has been shown to mitigate measurement error and minimize parameter bias (Glazerman, Levy & Myers, 2003; Steiner, Cook, & Shadish, 2011). A credibility test of the PSA design uses a continuation of the RCT treatment during the PSA study period to compare the PSA estimate on RCT treatment schools with the RCT estimate on these schools. Satisfying this credibility test implies satisfying the RCM requirement of unconfounded assignment. Equality of the two estimates for reading achievement suggests a credible PSA for reading, whereas for math achievement, the PSA was not found to be credible. The findings from this study suggested, in contrast to the RCT, that reading achievement was promoted by CareerStart. This study contributes to the larger scholarly work on quasi-experimental replication of RCTs in a study population.

CHAPTER 1. THE CONTEMPORANEOUS INFLUENCE OF CLASSROOM PEER ASSIGNMENT AND STUDENT-TEACHER INTERACTIONS ON TEACHER VALUE-ADDED MODEL ESTIMATION

Research into the effects of schooling inputs on student learning suggest that teachers are the most important input, such that improving teacher effectiveness is a legitimate policy target to promote greater student achievement and learning (Gordon, Kane & Staiger, 2006; Nye, Konstantopolous & Hedges, 2004; Rockoff, 2004; Rowan, Correnti & Miller, 2002). Timely and precise measures of teacher effectiveness must therefore be made available to state, district, and school administrators, and a number of different approaches have been proposed and used, including teacher (Rowan, Correnti & Miller, 2002) or student (Kahle, Meece & Scantlebury, 2000; Koth, Bradshaw & Leaf, 2008) survey data; and observation of teaching practice (Pianta et al., 2008; Schochet, 2011).

Presently, the most widely recommended approach for measuring teacher effectiveness is to estimate teachers' *value added* (TVA) to student learning (Tekwe, et al. 2004). The TVA estimates the amount of learning that each teacher contributes to her students' knowledge as it accumulates over multiple years of schooling, adjusting for achievement in past years. Using student learning as a metric for teacher effectiveness presents a substantial measurement challenge because in addition to students' existing skill and knowledge set, there are many time-varying or contemporaneous inputs to student learning that may confound measurement, including factors related to school (*e.g.*, Cook, MacCoun, Muschkin & Vigdor, 2008; Hanushek ,1996), home, family and neighborhood (*e.g.*, Bronfenbrenner, 1989; Heckman, 2008), peers (*e.g.*, Borman et al., 2004); and human

development (*e.g.*, Juvonen, 2006). Value-added modeling (VAM) comprises multiple types of statistical and econometric methods that purport to control for these influences and estimate an unbiased teacher effect.

Race to the Top (RttT), a federal program, requires states receiving RttT funding to rank and compare teachers statewide using VAMs (Henry, Kershaw, Zulli & Smith, 2012). Despite this mandate, the approach is controversial and several challenges to the adequacy of VAMs have yet to be completely resolved. These include 1) inconclusive evidence on whether teacher value added (TVA) estimates sufficiently adjust for previous learning and contemporaneous confounders (Chetty, Friedman & Rockoff, 2011; Goldhaber & Chaplin, 2012; Rothstein, 2010); 2) year to year instability in value-added estimates for individual teachers (Sass, 2008); 3) disagreement on which VAMs perform best, the criteria and methods for determining relative performance, and the level of transparency of these models (*e.g.*, Amrein-Beardsley & Collins, 2012; Guarino, Reckase & Wooldridge, 2012; Henry, Rose & Lauen, in review; Hill, Kapitula & Umland, 2011); and 4) that these methods may proscribe a limited view of teachers' knowledge and skills (Hill, Rowan & Ball, 2005; Kennedy, 2010).

In this study, I focus on the role of multiple classroom effects in the estimation of TVAs used to evaluate teachers, and in particular, their potential to induce bias. Classroom effects consist of the combined influence of non-random assignment of students to teachers and interactions among students and between students and teachers. In a causal interpretation of TVAs, these contextual effects are not viewed as characteristics of the teacher but instead as disturbances that bias teacher effect estimation (Reardon & Raudenbush, 2009; Rubin, Stuart & Zanutto, 2004). I demonstrate that estimating unbiased teacher effects in the

presence of these classroom effects disturbances may not be possible. In this study, I examine the robustness of four VAMs to two types of classroom disturbances, those based on non-random assignment, and those based on student-teacher and student-student interactions. I use two statistical and two practical criteria as the basis for comparing the VAMs. I then propose and examine three potential disturbance-reducing strategies.

In the following sections, I first discuss the importance of teachers and teacher evaluation, the role that VAMs are proposed to play, and the challenges associated with building a teacher evaluation system around student test scores and VAMs. I then discuss a causal framework, the Rubin Causal Model (RCM), which helps to clarify the methodological and conceptual challenges of this effort. Following this, I discuss the extant research on classroom effects and their role in TVA estimation. I then describe the methods and results of a study in which I examine the role of classroom effects in biasing teacher estimates; compare the performance of a set of typical value-added models in addressing these problems; and look at a number of potential analytic solutions. In doing so I model two simulations, each one designed to examine the failure of an assumption of the RCM. These simulations address one of the most challenging aspects associated with validating TVAs, that they are unobserved. By simulating data, I am able to create the true teacher estimate, but the stylized data are a limitation. I comment on this limitation in the discussion. In the discussion section, I also comment on the implications of this study for the use of VAMs in teacher evaluation systems; comment on the implications for the use of simulation in studying VAMs for teacher evaluation systems; and I raise questions about certain facets of teaching, such as those that emerge from classroom interactions, that may be evidence of

greater effectiveness but are discarded to meet the assumptions needed to obtain unbiased TVAs.

The Role of Quality Teaching in Student Achievement

Over the past three decades of research into the effectiveness of educational inputs on student achievement, scholars in diverse fields have examined the influence of contemporaneous influences on learning including home and family factors (*e.g.*, Coleman et al., 1966; Heckman, 2008; Todd & Wolpin, 2007), school resources (*e.g.*, Greenwald, Hedges, and Laine, 1996a; 1996b; Hanushek, 1981; 1996; Hedges et al., 1994), a variety of school organizational, cultural and structural factors (*e.g.*, Carnoy et al., 2006; Cook, MacCoun, Muschkin & Vigdor, 2008; Ladd, 2002), peers and peer diversity (*e.g.*, Borman et al., 2004; Juvonen, 2006; Woolley & Bowen, 2007), neighborhood and community influences (Bowen, Bowen & Ware, 2002), teaching practices (*e.g.*, Rowan, Correnti & Miller, 2002), teacher preparation (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2008; Goldhaber & Brewer, 2000; Harris & Sass, 2011; Smith, Desimone & Ueno, 2005), and human development (*e.g.*, Ames, 1992; Eccles & Wigfield, 2002; Juvonen, 2006).

The effort to rigorously and fairly evaluating teachers emerges partly out of recent studies that suggest the relative importance of teachers to student learning. These studies cross disciplinary boundaries and use a variety of designs and methods, and as a result the findings are widely regarded as robust. For example, using data from the Tennessee STAR experiment (of the effect of small class sizes on student achievement) and a variance decomposition method, Nye, Konstantopolous and Hedges (2004) found that teacher effects for students from kindergarten through third grade range from .12 to .135 for math achievement and .066 to .074 for reading achievement, which constituted about one-third a

standard deviation in math and one-half a standard deviation in reading between teachers at the 25th and 75th percentiles. Rockoff (2004) used teacher fixed effects models to show that teachers promoted higher student achievement in reading vocabulary (teacher fixed effect joint significance F = 4.43); reading comprehension (F = 2.74); math computation (F = 3.72), and math concepts (F = 5.30). Rowan, Correnti & Miller (2002) used student repeated measures growth curves to demonstrate that the classroom accounted for more than 60% of the variation in student reading test score growth, and 52-72% of the variance in math test score growth. Gordon, Kain and Staiger (2006) found a 10-percentile difference in achievement between students having teachers in the top and bottom quartiles of the distribution of teacher effectiveness.

The interest in teacher effectiveness or quality also represents a realignment of public priorities for education, representing the dynamic relationship between two often-competing values (Labaree, 1997; Sergiovanni et al., 1999), *equity*, our preference for equal opportunity and *quality*, the goal of ensuring that the education system helps children become informed, capable adults. After the publication of *A Nation at Risk* in 1983 (U.S. Department of Education), which argued for greater quality in public education, legislators in some states, such as North Carolina, implemented statewide assessment standards and incentives in the forms of rewards and sanctions for performance. At the federal level, reauthorizations of the Elementary and Secondary Education Act (in the form of No Child Left Behind legislation) have also addressed the quality deficit framed in terms of equity, such as requiring that students in a wide variety of demographic categories, such as special education or English language learners, meet pre-determined adequate yearly progress (AYP) standards. Further reflecting a convergence of quality and equity values, North Carolina's Leandro decision,

which ruled that several local education agencies (LEAs) did not provide an equal education for all students, helped to refocus educational priorities on equity, though in this case, framed in terms of quality as the outputs of schooling rather than in terms of funding or inputs.

Evidence suggests that interventions that improve teacher effectiveness may promote greater student achievement. For example, high stakes accountability consisting of rewards or sanctions relies largely on market forces to encourage ineffective teachers to leave the profession (*e.g.*, Gordon, Kain & Staiger, 2006), or to compensate highly effective teachers with performance pay bonuses (*e.g.*, Springer et al., 2010). Such methods are summative; alternatively, formative methods have been proposed (Amrein-Beardsley, 2008). In-service or professional development (PD) training programs may provide teachers with the knowledge and tools they need to become more effective (*e.g.*, Grossman et al., 2000; Woolley et al., in press) and may address deficits resulting from lack of experience or less effective pre-service training (Henry, Bastian & Fortner, 2011).

For education policymakers to act on these policies, precise measures of teacher effectiveness must be available. The typical evaluation system has, up to this point, consisted largely of principal observations of teachers at specified intervals during the school year, a method that is being subjected to greater rigor (Pianta et al., 2008). Research suggests a number of potential complements to the observational method. These include self-reported survey data obtained from teachers (*e.g.*, Hill, Rowan & Ball, 2005; Rowan, Correnti & Miller, 2002); teacher logs of their daily practices (*e.g.*, Hill, Rowan & Ball, 2005); survey data obtained from students as raters of their teachers (*e.g.*, Koth, Bradshaw & Leaf, 2008; Kahle, Meece & Scantlebury, 2000; Rose, Woolley, Orthner, Akos & Jones-Sanpei, 2012), a

practice that has long been used in the higher education system (*e.g.*, Marsh, Overall & Kesler, 1979).

Although a number of measures of effective teaching have been proposed, the focus on outputs, combined with the availability of student standardized achievement test score data, has led scholars and policymakers to promote evaluation systems based on students' accumulation of knowledge as measured by these test scores (Harris, 2009; Sanders, Saxton & Horn, 1997).

Teacher Evaluation using Value-Added Models

Value-added methodology comes from economic literature on production functions that describe industrial processes combining an input with technology to produce incremental value in a product. In the case of the *education production function*, the technology consists of schooling inputs such as teaching practice, the input is existing student knowledge, and the output is new student knowledge (Todd & Wolpin, 2003). A typical education production function (Todd & Wolpin, 2007) represents learning as a cumulative function with the following form:

$$(1) \qquad Y_{ihw} = Y_w(\eta_{ih}(w), \zeta_{ih0})$$

Y_{iha} is achievement for student *i* in household *h* in period *w*; Y_w is the achievement function evaluated in period *w*, with learning inputs up to period *w*, $\eta_{ih}(w)$, and intelligence, ζ_{ih0}. The education production function is usually given an additive and linear functional form as follows:

$$(2) \qquad Y_{ihw} = \eta_{ihw}\pi_w + \eta_{ihw-1}\pi_{w-1} + \ldots \eta_{ih0}\pi_0 + \pi_\zeta\zeta_{ih0} + \varphi_{ihw}\gamma_w + \varphi_{ihw-1}\gamma_{w-1} + \varphi_{ih0}\gamma_0 + e_{ihw}$$

The inputs in η_{ih} are represented here as observed inputs η_{ihw} and unobserved inputs φ_{ihw}, each corresponding to periods w = {0, … w}; e_{ihw} represents measurement error. Each observed and unobserved input is shown to contribute a different level to overall learning depending on how many years away from the current period the input was applied.

Not all education and education policy scholars see learning in such starkly mechanical terms and many reject the industrial analogy to the process of learning suggested by economic theories and the education production function (*e.g.*, Rutkowski, Rutkowski & Langfeldt, 2012). Nevertheless, value-added methods are now used widely as a means of assessing teacher effectiveness in education research and program evaluation (*e.g.*, Harris & Sass, 2011; Hill, Rowan & Ball, 2005). Because production functions are used to direct resources to industrial processes, an analogous use of the education production function is to direct resources to the learning process, in the form of policy interventions with teachers (Gordon, Kane & Staiger, 2006). By holding teachers to standards using *outputs* in this manner, policymakers would be moving away from, or at least offering an alternative to, standards based on inputs in the form of educational and credentialing requirements and principals' subjective observations (Harris, 2009).

Under a teacher evaluation system that uses VAMs as a source of information about teacher performance, teachers would be evaluated partly on the basis of their performance in what is widely viewed as their primary task: the educational progress that their students are expected to achieve during the course of the school year in which they are assigned to each other. In the cumulative-additive education production function of model 2, the teacher at age a is a school-specific portion of the *unobserved* input $\varphi_{iha}c_1$. A value-added model of cumulative student learning, developed from the cumulative-additive form of the education

production function distinguishes inputs in previous periods from contemporaneous inputs, and among contemporaneous inputs further distinguishes between student-specific and within-school inputs such as teacher, classroom and school, as follows:

$$(3) \qquad Y_{ijw} = \beta_Y Y_{ijw-1} + \beta_\mu \mu_{ijw} + T_{jw} + S_{sw} + C_{cw} + e_{ijw}$$

Relative to model 2, contemporaneous school inputs are given by teacher inputs T_{jw} for teacher *j* in period *w*, classroom inputs C_{cw} for classroom *c* in period *w*, and school inputs S_{sw} for school *s* in period *w*; μ_{ijw} represents contemporaneous student predictors of learning; and all inputs prior to period w are subsumed into Y_{ijw-1}. Any unobserved predictors of learning must be absorbed either by Y_{ijw-1} or the measurement error e_{ijw}, implying that the two components may be correlated.

Methods used to estimate T_{jw} as shown in model 3 include multilevel models for identifying the proportion of variance in test scores that can be attributed to teachers and students (Henry, Rose & Lauen, in press; Raudenbush, 2004); covariate adjustment to attribute portions of student variance in test scores to measured inputs to learning (Ballou, Sanders & Wright, 2004); econometric methods such as fixed effects that rely on year-to-year differencing in student achievement to completely eliminate other fixed inputs to learning (Rothstein, 2010); instrumental variables that adjust for autocorrelation in learning (Arellano & Bond, 1991); or combinations of these approaches (Guarino, Reckase & Wooldridge, 2012). Variations on model 3 are used throughout the literature on TVA estimation, some of which treat T_{jw} and C_{cw} as one, or leave S_{sw} out of the model altogether. The objective of using the model in this context is to sufficiently specify model 3 such that T_{jw}, which is not usually observed, can be estimated as the value added to learning that is leftover after accounting for all other contemporaneous sources, as well as prior knowledge:

$$(4) \qquad T_{jw} = Y_{ijw} - (\beta_Y Y_{ijw-1} + \beta_\mu \mu_{ijw} + S_{sw} + C_{cw} + e_{ijw})$$

There are two challenges to estimating TVAs from model 4. The first is that contemporaneous classroom and school inputs are generally not measured directly in administrative data (Raudenbush, 2004), and as a result, they are potentially taken up by the error, or absorbed into the teacher effect T_{ijw} to the extent that C_{cw} and S_{sw} are correlated with T_{ijw}. The second is that because T_{ijw} is estimated from leftover variation after taking measured inputs into account, T_{ijw} must be distinguished from the error, e_{ijw}, which captures any unmeasured contemporaneous inputs to learning. A more accurate depiction of the challenge represented by model 4 for any teacher *j* is given by 4':

$$(4') \qquad T_{jw} + e_{ijw} = Y_{ijw} - (\beta_Y Y_{ijw-1} + \beta_\mu \mu_{ijw} + S_{sw} + C_{cw})$$

In situations where C_{cw} and S_{sw} are not measured, the effects of these inputs would be subsumed into the terms on the left hand side. In research settings where population averages are estimated, the expected value of the error is assumed to be zero. However, in teacher evaluation settings, assuming that the expected value of the error for an individual teacher is zero is a much stronger assumption that may not be possible in the context of the relationships between the teacher effect and the other inputs. As this study and others demonstrate, these are both significant challenges.

Advantages over Existing Approaches

Measuring teacher effectiveness via outcomes may address limitations inherent in other approaches that have been used, including using teacher certification or preparation information; using survey data collected from teachers or students; using observational data; and using unadjusted student achievement data.

Certification & Training. Studies have demonstrated only a weak link between student achievement and structural characteristics of teacher preparation, including board certification (Darling-Hammond, Berry & Thoreson, 2001; Golhaber & Brewer, 2000), the number of content area courses (Boyd, Grossman, Lankford, Loeb & Wyckoff, 2009), PRAXIS II exam scores (Xu, Hannaway & Taylor, 2011), college entrance exam scores (Harris & Sass, 2011), route of preparation to teaching, such as through traditional preparation (Henry, Bastian & Fortner, 2011) or lateral entry such as Teach for America (Glazerman, Mayer & Decker, 2006; Xu, Hannaway & Taylor, 2011), or the number of in-service hours of training (Harris & Sass, 2011). Experience, instead, is a much more important predictor of teacher effectiveness (Gordon, Kane & Staiger, 2006; Henry, Bastian & Fortner, 2011). Economic theory suggests that certification and training programs are actually barriers to the efficient flow of human capital into and out of the teaching profession (Rutkowski, Rutkowski & Langfeldt, 2012; Xu, Hannaway & Taylor, 2011).

Surveys. Survey data can be obtained from teachers (Rowan, Correnti & Miller, 2002) in the form of self-reports of the practices that they use, but these may be subjected to some desirability bias. Student survey data on teaching practice may solve the desirability bias problem and yield more reliable multiple ratings per teacher (Guo & Hussey, 1999; Kahle, Meece & Scantlebury, 2000; Koth, Bradshaw & Leaf, 2008), but may be affected by recall biases and subjective differences based on the same home and peer experiences that influence learning (Woolley & Rose, unpublished draft).

Observations. Observational data of teaching practice have typically been collected by principals, peers, or independent observers (Hill, Charalambous & Kraft, 2012; Pianta et al., 2008). The observational method suffers from low reliability (Schochet, 2011), as well as

subjectivity of the observer, the choice of instrument, and the number and types of lessons to observe (Hill, Charalambous & Kraft, 2012). The use of rubrics and rigorous training in the use of these rubrics can reduce both the influence of subjectivity and the unreliability of observational measures (Pianta et al., 2008). Principal observations have been shown to be equivalent to VAMs in identifying the worst and best performing teachers (Jacob & Lefgren, 2008).

Unadjusted Student Achievement. District administrators and principals may already use educational outputs, in the form of student test score data, to evaluate teachers, but these educators may be using less sophisticated and untrustworthy methods, such as average student performance on end-of-grade tests (Andrejko, 2004). Because average student performance is a function not only of a teacher's effectiveness in a given year but also of factors unrelated to that teacher, such as the effects of previous teachers and contemporaneous and historical factors related to neighborhood, home, peers, and family, a measure of teacher performance from average student achievement may unfairly reward or penalize teachers for factors beyond their control (Guarino, Reckase & Wooldridge, 2012). The validity of average student performance as a measure of teacher effectiveness depends on whether teachers are effectively randomly assigned to students. If random assignment is the norm, then any factors not due to the teacher can be assumed to be randomly distributed and average student performance is adequate (Guarino, Reckase & Wooldridge, 2012). The quantitative evidence suggests that this is not the case, because the means of assignment does not effectively randomize over students' or teachers' abilities (Clotfelter, Ladd & Vigdor, 2010, 2005; Horng, 2009).

VAMS, alternatively, are hypothesized to account for these factors, rendering the student-teacher assignment conditionally random and allowing the TVA to be interpreted as the teacher's effectiveness in adding to learning. VAMs represent some promise for an objective, outcome-based evaluation of teacher effectiveness, but as the evidence shows, there are important concerns, some of which have not been addressed.

Concerns About VAMs as Evaluation Tools

Teacher VAMs have been successfully applied in education production function research to identify causal inputs to schooling (*e.g.*, Hanushek, 1986; Harris & Sass, 2011). It is this success that partly explains the interest in translating these research models to the practice of evaluating teachers (Sanders, Saxton & Horn, 1997). But some aspects of these research methods represent challenges to this effort. For example, in research that is typically concerned with population averages, an error of 5% in identifying a chance input as systematic, when that input is a population parameter representing an average tendency in a population, is acceptable. It implies that once out of every 20 similar studies we can expect to commit an error and demonstrate an effect where there is none. In evaluating individual teachers, which implies a focus on unit-specific estimation, a 5% rate of error may not be acceptable. Lower rates than this (*e.g.*, 1.2% error in properly identifying a teacher as ineffective) can translate to a high number of teachers on a statewide basis.

A second concern about applying these methods has to do with the transparency of the models and the buy-in from the stakeholders being evaluated. Certain VAMs are very sophisticated and as a result teachers, administrators, and parents may not view VAMs as transparent, and therefore decisions about model form or specification are not subjected to the input of these stakeholders (Hill, 2009; Amrein-Beardsley, 2008). Although those in

favor and those against VAMs represent a wide variety of disciplinary backgrounds, VAMs are viewed more favorably by education policy researchers (*e.g.*, Harris, 2009), and not viewed as favorably by scholars involved in educator training (*e.g.*, Amrein-Beardsley, 2008). Because they may feel that these systems did not take their concerns, views, or expertise into account, some teachers and administrators may reject the findings of these evaluation systems, particularly if these findings contradict their own subjective experiences. One mixed method study reported that teachers who have been evaluated using VAMs indicate that their VAM scores have varied from year to year despite invariance in their teaching practice, implying that context varied and was insufficiently controlled for by the VAMs (Amrein-Beardsley & Collins, 2012).

These concerns are not trivial. Educators may rebel against outside influence in the education process. For example, in January 2013, teachers in a Seattle school refused to administer standardized achievement tests that would be used to evaluate teachers, citing concerns about whether the test measures learning, and consequently whether it is an appropriate measure for teacher evaluation (Scott, 2013). In the week-long strike by the Chicago Teachers Union in 2012, the teacher evaluation system played a significant role, with the union opposed to efforts to increase the role of student standardized test scores in teacher evaluation; the union prevailed, and in the new system only 30% of teachers' evaluations (the state minimum) will be based on test scores (Pearson, 2013).

To use standardized end-of-grade tests for measuring student learning as markers of teacher effectiveness, evidence of their validity as measures of the teacher's contribution to student learning, and their reliability from year to year in measuring this contribution with minimum error is needed. I now turn to discuss this issue.

Measurement Validity and Reliability

A teacher evaluation system based on student test scores as measures of student learning must first demonstrate validity and reliability in measuring student learning. Validity of the test concerns systematic error in measurement, or that it measures the construct it purports to measure, in the context of its intended use (Hill, Kapitula & Umland, 2011; National Research Council, 1998). Thus, a seventh grade reading test is a valid measure of students' reading skills for advancement to eighth grade if it accurately assesses the skills and knowledge needed to read at the seventh grade level as a prerequisite for eighth grade reading. Alternatively, if it assessed the skills needed to read at the sixth grade level, it would not be valid; nor would it be valid if it assessed only seventh grade vocabulary rather than the full complement of skills needed to read at this level, or if it assessed seventh grade reading and biology (National Research Council, 1998). Validity would also be called into question if instruction did not adequately prepare students for eighth grade reading, such as through test preparation or coaching to yield higher seventh grade results without commensurate gains in knowledge (Koretz, 2008).

Reliability concerns whether a specified form of the test measures this domain with minimal random error includes internal consistency reliability (reliability across subsets of items), test-retest reliability (reliability over repeated administrations of the same set of items) and equivalent-form reliability (reliability of instruments with different sets of items; National Research Council, 1998). Failures of these forms of reliability represent challenges for making inferences about learning; for example, if an instrument fails to show test-retest reliability, it may not be possible to infer that an improvement in a test score implies an increase in knowledge.

Studies question whether the domain of content and skills students need to learn can be accurately measured and that the items that are prepared measure these content and skills (Koretz, 2002; Linn, 2000; Reckase, 2004); whether the tests can measure higher order thinking skills (Kennedy, 1999); whether test content narrows the curriculum (Linn, 2000); whether the tests assess learning rather than test preparation, given the influence that coaching and test preparation may have on the manipulation of student test scores (Koretz, 2008); whether test scores correlate with known predictors of student learning and achievement (Pianta et al., 2008); and whether standardized tests have test-retest reliability (Koretz, 2002).

At the level of student learning, standardized test scores have been called into question. Extrapolating further from individual students' scores to aggregates of students' scores at the teacher level introduces further challenges to validity and reliability. The amount of learning experienced in the new period as measured by a difference between achievement in the current period and the previous as represented by model 4, even leaving aside the other contemporaneous inputs to learning, requires an assumption about the persistence of previous periods' learning and teachers' effects (Lockwood, McCaffrey, Mariano, & Setodji, 2007; Raudenbush, 2004). Typically, full or partial persistence is assumed, implying a fixed persistence parameter representing the proportion of previous years' learning remaining at the time new learning is assessed (*e.g.*, Guarino, Reckase & Wooldridge, 2012). Lockwood et al. (2007) used Bayesian methods to estimate a VAM with an empirically derived a persistence parameter.

The question of persistence concerns associating an amount of learning with two or more teachers over multiple years. The dilemma of associating a specified amount of

learning with multiple teachers can occur within a single year. Different types of instructional methods involve multiple teachers, including those that have more than one teacher or a teacher and teaching assistants in a classroom; those that involve outside tutoring and assistance; and those in which students change teacher assignments during the school year (Valli, Croninger & Walters, 2007). In these cases, apportioning learning gains to any teacher may be a challenge.

These are substantial measurement challenges, but viewing TVA estimation strictly as a measurement problem through the lens of validity and reliability overlooks a key feature of VAM methodology. Because VAM-based evaluation systems are intended to attribute aggregate changes in *student* achievement to the *teacher*, there is an implied causal relationship in the teacher effect measure (Rubin, Stuart, & Zanutto, 2004).

Attributing Student Test Score Changes to Teachers

Causal processes are typically absent from conventional measurement efforts. For example, measuring student engagement using a set of items on a self-report instrument implies that the reporter and the subject are one and the same, and thus no causality is implied. Similarly, asking a student to report on his perception of his teacher's practices does not imply a causal relationship. Although the reporter and the subject are not the same, the report is a subjective observation by the student reporter of the teacher and no causal process is implied. For a teacher effectiveness measure to be equivalent in design, a student would have to be asked to report back on their teacher's effectiveness in promoting their learning during the school year, a subjective measure. An analogous subjective measure would involve asking a student to describe how much of the material she learned in a given year was due to her teacher, which would be affected by perceptive biases. However, the TVA

uses independent assessments of learning progress that are generally viewed as objective rather than subjective. The measure implies a relationship between the teachers' practices and the level of performance demonstrated by the student on the learning assessment. *If certain assumptions are met,* the TVA can be viewed as causal estimates of the effects of teachers' practices on each individual student. If they are not, then it is unclear whether the TVA can actually be fully attributed to the teacher, which raises questions about whether TVAs provide useful teacher evaluation information, and how such information is subsequently put to use.

Viewing the measurement problem through the lens of this causal relationship helps to clarify the challenges of validity and reliability as challenges of establishing that the teacher effect estimate is a quantitative measure of factors attributable strictly to the teacher, including training, knowledge and teaching practices (validity), and that accordingly teacher ranking should not change unpredictably from year to year (reliability). A causal model provides a framework for demonstrating how typical value added models may violate these assumptions in the presence of classroom effects. The causal model that I will use is the Rubin causal model (RCM), also known as the potential outcomes model (Holland, 1986). I focus on classroom effects, a contemporaneous input to learning, as the central challenge described by the model.

The Rubin Causal Model: Assumptions and Violations

In the Rubin causal model (RCM), the definition of the causal estimand for a specified teacher's effect depends on the counterfactual condition—the conditions that prevail in the absence of the specific cause—that is, all other teachers. Formally, assume that the outcome for student i (with I = 1,…N) under treatment condition j is Y_{ij}, with J possible

teachers in the population of teachers (*e.g.*, statewide for a statewide evaluation system) as treatments, a many-valued treatment with the potential outcomes represented by a matrix of *N* students by *J* teachers (Morgan & Winship, 2007; Reardon & Raudenbush, 2009). The student has one potential outcome Y_{ij} under each teacher, where $Y_{ij}$is usually academic achievement. In what is known as the fundamental problem of causal inference (Holland, 1986), only one of these potential outcomes can be realized given that each student can only be assigned to one teacher at a time. Instead, students assigned to one condition are compared to students assigned to the other conditions. This is done using a function of the distributions of performance of students assigned to treatment j and the students under the other condition or conditions. Typically this function is the average treatment effect for treatment j (ATE_j), estimated as the difference between the average performance of students observed under assignment to treatment j and the average performance of those assigned to (for example) the average teacher, labeled $.j$ (not j).

There are three assumptions in the RCM—manipulability, stable unit treatment value, and unconfoundedness (Reardon and Raudenbush, 2009). Manipulability implies that each student has a potential outcome under each teacher in the population (manipulability). The stable unit treatment value assumption (SUTVA) implies potential outcome under each teacher is independent of the assignment of peers. Unconfoundedness implies that student-teacher assignment is independent of students' potential outcomes. When satisfied, these assumptions make it possible to infer the treatment effect as causal despite the fundamental problem of causal inference that only one of *J* potential outcomes can be realized. Violations of any of these assumptions may occur. Typically, the manipulability assumption is understood to be violated, because school attendance zones are determined largely by

income, and therefore it is not plausible to assume any student can be assigned to any school or teacher (Beller & Hout, 2006; Reardon & Raudenbush, 2009). Despite the implausibility of manipulability happening in practice, it is not impossible to imagine, and in fact has happened in randomized studies such as the Tennessee STAR class size experiment (Finn & Achilles, 1990). Therefore, the assumption is often seen as harmless (Reardon & Raudenbush, 2009). Violations of SUTVA and confoundedness as a result of classroom and peer effects are the focus of the present study.

Stable Unit Treatment Value Assumption

SUTVA implies that the teacher's effect on any student does not vary according to the presence of the other students in that teacher's classroom (Reardon & Raudenbush, 2009; Rubin, Stuart & Zanutto, 2004). The presence of interactions between students in classroom learning that produce net differences in student achievement may lead to an increment to (or decrement from) the teacher effect that would make it more difficult to obtain an unbiased estimate (Henry & Rickman, 2007). Formally,

$$\text{(i) } Y_{ij}(AZ) = Y_{ij}$$

AZ is an *N* x *J* matrix of *ij* elements recording assignment of students to teachers, with *ij* = 1 if *i* is assigned to *j* and *ij* = 0 otherwise. The statement above makes it explicit that Y_{ij} is invariant to all *Z* permutations of *A*, a vector indicating each student's assignment to treatment.

Unconfoundedness

Unconfoundedness, also known as ignorability or exogeneity, implies that each student's assignment to treatment – that is, their assignment *A* to a specific teacher– is independent of their potential outcome under that teacher (Morgan & Winship, 2007):

$$\text{(ii) } Y_{ij} \perp A$$

The formal statement *ii* is the strictest version of this assumption, strong ignorability. A weaker form conditions on student background factors, x:

$$\text{(ii') } Y_{ij} \perp A \mid X = x$$

Many VAMs ignore covariates and those that account for covariates handle them differently either by explicitly specifying them or by using fixed effects to eliminate their influence (*e.g.*, Ballou, Sanders & Wright, 2004). The strong form is violated under any associations with measured or unmeasured covariates. The weak form is violated only in the presence of association between omitted variables – unmeasured inputs to learning– with each student's potential outcomes, and each student's assignment to teachers (Reardon & Raudenbush, 2009). The weak form is violated because typically, large-scale administrative data files do not contain measures of all of the inputs to learning (including, for example, the availability of reading material in the home). Because groups of students are not usually randomly assigned to teachers their influences cannot be safely disregarded. The unconfoundedness assumption establishes a "boundary" on the extent of this problem however, by clarifying that these unmeasured covariates of learning are confounders of the teacher effect only if these inputs are also associated with treatment (student-teacher) assignment, reducing the burden for inferring causality.

Unaccounted-for classroom effects can both lead to violations of SUTVA and confoundedness, and thus they both reduce to measurement problems. SUTVA violations can emerge within the classroom as a consequence of interactions between groups of students and between students and the teacher. Further, they can be related to non-random assignment patterns (*e.g.*, Henry & Rickman, 2007; Mashburn, Justice, Downer & Pianta, 2009).

SUTVA can occur even in scenarios where the treatment conditions are randomized. Non-random assignment leads to violations of the unconfoundedness assumption and may affect TVA estimation by failing to provide a fair baseline against which to evaluate the teacher, though some methods may be more robust to this violation than others. SUTVA and unconfoundedness violations caused by the classroom environment are the focus of the present study.

The Classroom Effect in the Context of the RCM

The classroom is the environment for teaching and the setting for practice, which is influenced or affected by teacher training and inherent characteristics of teachers, and student assignments to teachers. These factors subsequently give rise to student peer and teacher-student interactions. The classroom effect is an emergent property of the students and teacher that varies according to these factors (Kennedy, 2010; Mashburn, Justice, Downer & Pianta, 2009; Rothstein, 2010; Skinner & Belmont, 1993). *Aggregate pre-assignment characteristics* of students typically consist of the average ability or preparation of students assigned to a teacher of a certain effectiveness level (Rothstein, 2010). *Student peer interactions* produce learning effects (good or bad) beyond those directly from the teacher and have the potential to alter the teacher's perceived effectiveness (Reardon & Raudenbush, 2009). *Teacher-student interactions* are viewed as consisting of a unidirectional transfer of knowledge and skills from the teacher to the student, but evidence suggests that the teacher may alter his teaching practices as a result of these interactions, suggesting that different groups of students may lead the teacher to act differently and use different practices (Hamre et al., 2012). Hanushek, Kain, Markman and Rivkin (2003) classify these into exogenous (pre-assignment) and endogenous or behavioral (post-assignment). Aggregate pre-assignment

characteristics are typically thought of as a manifestation of confoundedness and the post-assignment factors thought of as violations of SUTVA that may arise as a result of this confoundedness or independent of it. I now turn to discuss these factors in detail.

Aggregate Pre-Assignment Characteristics

Non-random assignment, being widely viewed as the key problem that successful VAMs must address, has received most of the attention in the TVA literature. Thus far, the evidence is mixed (*e.g.*, Tekwe et al., 2004). Raudenbush (2004), for example, suggests that the causal question being answered by VAMs—about the difference between the effect of treatments (teachers)—cannot be answered without bias because factors confounding teacher effectiveness cannot be easily controlled for without random assignment.

In a landmark study, Rothstein (2010) demonstrated that teacher assignment in period *w* (*e.g.*, fifth grade) was associated with average test score gains in period *w-1* (*e.g.*, fourth grade) implying that three econometric VAMs (change scores on grade and classroom dummies; regression of gains on classroom indicators and lagged score; and a panel of multiple gain scores with student FE), despite their statistical sophistication, did not sufficiently address the issue of confoundedness due to unmeasured variables associated with teacher-student assignment. Rothstein's finding were replicated by Koedel & Betts (2009), who subsequently demonstrated that by using multiple years' data, some of the confoundedness could be alleviated. Rothstein's evidence for biasedness operated primarily through negatively correlated errors in measurement across consecutive grades.

Goldhaber and Chaplin (2012) showed that the Rothstein test can actually reject the null hypothesis of no association between *w-1* scores and *w* assignment when there is no confoundedness from omitted correlates of teacher assignment; that is, it could be completely

artifactual, implying that the Rothstein test is not a formal condition for confoundedness. They describe three conditions in which the test would artifactually be rejected. First, current teachers' could be conditionally associated with student achievement from two periods prior, which is often the most recent data available to principals to make assignment decisions as one period prior (*e.g.*, May) is not usually available at the start of the current period (*e.g.*, August; that is, there is a lag of time before the most recent test scores become available). Second, current teachers may be associated with previous teachers possibly due to tracking as students advance together from one teacher to the next. Third, the current teacher may be conditionally correlated with the previous period's error term—which represents the condition the Rothstein test was devised to detect. Goldhaber and Chaplin use a simulation study to show that all three manifestations can be detected by the test, despite the third being the only condition under which the teacher effect would be biased.

Chetty, Friedman and Rockoff (2011), alternatively, use a quasi-experiment and measured parent data to suggest that conditional on these parent variables, there was no confoundedness. The authors used matched tax and school district data on 2.5 million third through eighth grade children and their parents. TVAs were estimated using Empirical Bayes shrinkage using only the district data, and a fixed effect strategy that uses cross-school movements in teachers. Specifically, if the TVA was confounded by student background, and a teacher with a high observed TVA score moved from one school to another, the students in her new school would be expected to have lower test scores implying the TVA was due to her previous students. Using this strategy, the authors show that the parent data provides no additional information beyond what is provided by typical school data for identifying unconfounded estimates of teacher effects.

Student Peer Interactions

Scholars in diverse fields have demonstrated that peers exert an influence on each others' learning and that separating them from the teacher effect may be a challenge (McCaffrey et al., 2004). The idea that peers may influence learning goes back to Coleman et al. (1966). In contemporary education policy research, student peer interactions are not usually measured directly (an exception is Kinderman, 2007), and much research in this area continues to utilize predictors of peer interactions as measures of peer interactions. Several studies have demonstrated important findings about the efficacy of student peer interactions in promoting achievement and learning.

Although not all studies find significant peer effects (*e.g.*, Harris & Sass, 2006), studies in education policy research, on balance, suggest the importance of peer effects to learning. Lefgren (2004), using school tracking mechanisms as a natural experiment, demonstrated small but statistically significant positive peer effects. Specifically, in comparison to students in untracked settings, students in tracked settings would be expected to perform higher if peer effects promoted positive gains to learning. Both instrumental variables and fixed effects were used to control for omitted variables.

Zimmer and Toma (2000) used selected international public and private school data from five countries to demonstrate that peer effects were significant, and of greater magnitude for low-ability students as compared to high ability students. Peer effects were defined based entirely on existing ability as measured by the mean, squared mean and standard deviation of previous achievement of students in a classroom, as well as an interaction of this mean with students' own previous achievement; alternative forms included the proportion of high and low ability students. The class mean was significantly positively

associated with achievement; the squared mean suggested that gains declined as average classroom performance increased; higher variability translated to higher performance; and the interaction of previous peer and self achievement suggested that lower achieving peers benefited more. In the alternative form of the model using proportions of high and low peers, more high-achieving peers promoted lower achievement gains, while no effect was observed for the proportion of low achieving peers. Burke and Sass (2008), using peer fixed effects rather than functions of aggregate previous achievement, found similar small and positive effects but also note that peer and teacher quality may co-vary.

Hanushek, Kain, Markman and Rivkin (2003) defined two classes of peer effects, endogenous or behavioral effects and exogenous, predetermined or contextual effects that are not affected by in-class behavior, and use fixed-effects value-added specifications to control for unobserved heterogeneity in their examination of peer effects in three cohorts of Texas elementary school students. Using a measurement strategy similar to Zimmer and Toma, they demonstrate using a variety of model specifications significant effects of average prior achievement, standard deviation in this measure, and the proportion eligible for reduced price lunches in predicting math achievement levels and test score gains, with the most robust finding coming from average previous achievement.

Henry & Rickman (2007) used a series of robust value-added specifications of preschool skills to demonstrate that peer effects were significant predictors of cognitive skills, pre-reading ability and expressive language skills conditioned on family and school characteristics and child skill level and enrollment among preschool children in Georgia. Further, they demonstrate that classroom composition, teacher motivation, and time on discipline do not explain the transmission of these effects. Peer effects were measured as

class averages of a composite of pretreatment measures, including cognitive skills, pre-reading skills, expressive language skills and number of basic skills mastered.

Studies in educational psychology support these findings and provide deeper insight into the mechanisms by which peers influence learning. For example, friends may provide direct help in the form of assisting with homework and problem solving (Juvonen, 2006). Alternatively, the mechanisms may be psychological or emotional. For example, students' values and behaviors may be moderated by their perceptions of their peers' values, including reinforcing either negative or positive behaviors, a trend that is stronger in older (*e.g.*, middle school) students than in younger students (Juvonen, 2006).

Kinderman (2007) examined whether student engagement as measured over multiple periods was associated with characteristics of early-schooling natural peer groups, where by natural the author meant the peers that students interact with most frequently as revealed by socio-cognitive mapping, rather than students' subjective views of friends or acquaintances. Socio-cognitive maps were developed from surveys asking students to report back on interactions with peers as identified by name. Subsequent structural equation models examined the relationships between these groupings and engagement and demonstrated that students' changes in engagement during the school year were significantly predicted by their peer group affiliations, though the magnitude was small (two% of variance). Affiliating with high-engagement peers predicted same-level or increased engagement, whereas those in networks of less-engaged peers reported decreases.

Teacher-Student Interactions

Teacher-student interactions are the day-to-day social and instructional exchanges between teachers and students (Hamre et al., 2012). Teachers set the classroom environment,

including norms and rules (Ryan & Patrick, 2001); academic press (Hamre et al., 2012; Juvonen, 2006; Lee & Smith, 1999); motivational approaches including intrinsic and extrinsic motivational strategies (*e.g.*, Lepper, Corpus & Iyengar, 2005), goal structures (Ames, 1992; Anderman et al., 2010; Pintrich, 2000); classroom organization (Cameron, Connor, Morrison & Jewkes, 2008); classroom management (Kagan, 1992); and emotional support (Hamre et al., 2012; Lee & Smith, 1999). Further, pedagogical content knowledge, or the knowledge needed to teach the content (Shulman, 1987; Van Driel & Berry, 2012), requires knowledge about how students learn the subjects they are being taught (*e.g.*, Hill et al., 2008). All of these involve interactions with students.

Typically these manifestations of teacher-student interactions are seen as unidirectional from teachers to students and may be an inherent part of the teacher's effectiveness. However, research has demonstrated that these interactions are bi-directional and teachers' approaches to individual students are informed by those students' behaviors, attitudes, and cognition, implying that they may be promoted or hindered by different groupings of students. For example, teachers may treat students differently on account of race/ethnicity or gender (Roeser, Eccles & Sameroff, 2000), or may take a more favorable attitude towards students who are highly academically motivated (Juvonen, 2006) and adjust their motivational practices to meet the motivation demonstrated by their students (Jang, Kim & Reeve, 2012). As the following paragraphs show, several studies use structural equation models to allow for bidirectional and reciprocal effects between these variables.

Engagement and motivation have been shown to predict the quality of the teacher-student relationship. Skinner and Belmont (1993) show that higher teacher involvement, structure and autonomy support was predicted by children's motivation during the school

year, indicating reciprocal effects of teachers' early-year interest in students' motivation and suggesting that if students are undermotivated, during the course of the year their motivation will decline. In a study of self-determination and autonomy in secondary school classrooms, student engagement was found to predict teacher support (Van Ryzin, Gravealy & Roseth, 2009). In a study of first grade children in a Texas school district over three years, student engagement in year 2 predicted teacher-student relationship quality in year 3 (Hughes, Luo, Kwok & Loyd, 2008).

Reciprocity has also been demonstrated in the relationship between non-cognitive student measures and teacher-student relationships. In a separate study on the children in the Texas study conducted at a later time, peer relatedness was shown to predict higher teacher-student relationship quality (Hughes & Chen, 2011). Antisocial conduct also has a reciprocal relationship with teacher practice, with aggressive student behavior increasing conflict with teachers, which in turn worsened student aggression (Stipek & Miles, 2008).

The Classroom Effect as a TVA Disturbance

The purpose of this study is to understand the role that these contemporaneous classroom effects may play in biasing TVA estimates of teacher effectiveness and the implications that this bias may have on teachers subjected to an evaluation system deployed in a typical state. The purpose of the evaluation system—the question that the evaluation system is designed to answer—is required to choose the counterfactual for each teacher's effect (Raudenbush, 2004). In the context of this study of the influence of classroom influences on TVA estimation, I limit this choice to two possible counterfactual conditions.

If the purpose of the teacher effect estimate is to understand whether the teacher is effective in a particular setting (*i.e.*, with a particular group of students), then classroom

effects may not violate SUTVA or confoundedness and the TVA may be interpreted as causal for that setting, peer group, and teacher. This implies comparing a teacher's performance in one setting with her performance in another setting. Formally, the teacher effect may be labeled T_{jk} (rather than *j*; where *k* indicates classroom), with a counterfactual of *j*, not *k* (*i.e.*, $T_{j.k}$) representing teacher *j*'s effect in settings other than *k*.

Effect T_{jk} would enable comparisons within a single teacher over time or classrooms, but would not be acceptable for a comparison of teacher *j* with other teachers (.*j*). However, the purpose of the typical teacher evaluation system is to compare between teachers irrespective of setting, such that teachers can be compared with other teachers and subsequently be rewarded, sanctioned, or provided professional development and coaching. In this case, confoundedness and SUTVA may be violated to the extent that teacher assignment is effectively non-random and there are student-student and student-teacher interactions. If such violations exist, then different treatment effects will emerge from different combinations of students having the same teacher and teachers will be rewarded or sanctioned for reasons other than teacher effectiveness, regardless of the actual assignment mechanism.

As the purpose of this study is to illustrate the influence of contemporaneous classroom effects on TVA estimation and the implications for an evaluation system of this type, I assume that the classroom effects from confoundedness and SUTVA are undesirable and should not be included in the TVA estimate. In the case of both SUTVA and unconfoundedness violations, positive associations between the classroom and teacher effects will produce stronger teacher effects, biasing the TVA upward, although negative associations will bias the TVA downward. VAMs that incorporate features that minimize

these biases are better from a policy perspective, benefitting the students that these teachers have as well as being fairer to the teachers themselves.

Proposed Solutions to the Classroom Effect Disturbance

Ideally, the factors underlying these violations are measured and properly specified in estimating the teacher effect. However, as noted above, many of these characteristics are simply unavailable in the data that are used for TVA estimates. The resilience of typical VAMs to classroom effect disturbances, and their resulting performance in estimating the true teacher effect in the presence of this disturbance, is not known. Barring that all of these models are completely resilient, solving the misattribution problem is essential to accurate teacher effect estimation, particularly if individual teacher rewards or sanctions are at stake. I propose three methods of reducing the effect of classroom effects disturbances on TVA estimation.

Multiple cohorts of students. One option is to use multiple cohorts of students, which increases the number of teaching settings in which the teacher is observed. This method has been shown to stabilize teacher effect estimates, making them less likely to fluctuate from year to year (Koedel & Betts, 2009). Consequently, this approach is, from a measurement standpoint, reliability-promoting because using multiple observations of the teacher reduces the amount of random noise picked up by the teacher estimate. I argue, however, that it may also be validity-promoting in that it may make clearer the distinctions between classroom and teacher effects by averaging over the effects of multiple classrooms. In elementary school classrooms, this requires observing the teacher over multiple years in the same grade level, as teachers usually teach the same students over the course of the day. It is not clear how many cohorts would be needed for the biasedness from the classroom

effect to become negligible. However, if the evaluation system is to provide timely information to principals, at most three to five years would be possible, as this is typically the length of time for new teachers to level off in their effectiveness and for less effective teachers to sort themselves out of the teacher labor market (Henry, Bastian & Fortner, 2011). This strategy applies to all of the models, and no modifications to the models are required.

Multilevel model with classroom level. The variance in the student outcome at the classroom level is between-student, within-teacher variance that if not properly accounted for will manifest on the teacher effect. One strategy to address the disturbance created by classroom effects is to use a multilevel framework with a classroom level or variance component added to estimate the classroom variance separately, preventing it from manifesting on the teacher variance (Raudenbush & Bryk, 2002; Schochet & Chiang, 2010). This strategy will only work if teachers are observed teaching the same subject and grade level to more than one cohort or class of students; otherwise, the classroom and teacher effect will be the same and the model will be over-identified. The baseline simulation satisfies this condition, having two cohorts of students. This is not a strategy that applies to all of the VAMs in this study but instead comprises a fifth VAM.

Classroom covariates. As an alternative to adding a variance component, I add a composite classroom covariate to the multilevel model to absorb the classroom effect such that it does not bias the TVA estimate. Including classroom covariates is a common strategy (*e.g.*, Harris & Sass, 2011; Rockoff, 2004). In these situations, the classroom covariates are often simple aggregates of student characteristics much like the methods used in Zimmer and Toma (2000). Although this approach is desirable for eliminating the classroom disturbance, model 4 (which shows that the teacher effect is estimated as the difference between student

outcome and the linear combination of all measured inputs) suggests that there may be two problems. First, this approach will only work to the extent that measured classroom variables explain all of the potential classroom effects. If residual unmeasured classroom effects remain after conditioning on these variables, they will be attributed to teacher effect. Second, the inclusion of the measured classroom variables will bias the teacher effect estimate to the extent that these variables are correlated with the teacher effect (Ballou, Sanders & Wright, 2004). Diagram 1.1 shows how shared variance is treated in a regression framework, and demonstrates that any shared variance between the teacher and the classroom effect will be absorbed by the variables entered for the classroom effects. The net effect of these two sources of bias on the teacher effect is unclear.

Objective and Research Questions

In two previous studies using these simulations, Henry, Rose and Lauen (under review) and Rose, Henry and Lauen (unpublished manuscript) tested nine VAMs, demonstrating that four of these models—a random effects model, a student fixed effects model, a hybrid model, and a pooled ordinary least squares model—yielded estimated teacher effects that were more similar to simulated true teacher effects than five other models (also see Guarino, Reckase & Wooldridge, 2012). Table 1.1 is a summary table of all nine models, including the four examined in this study. In this study, I focused on these four VAMs. I propose to answer the following questions. First, how much does misattribution of classroom variances and fixed effects bias the teacher effects? Second, using a set of empirical and practical criteria, how do four typical VAMs compare to a true simulated teacher effect? Third, how does this comparison change as the proportion of variance attributed to the classroom increase (manifestation of SUTVA violations in the absence of

confounders) or as the correlation between the classroom fixed effect and teacher fixed effect increases (manifestation of confoundedness)? Fourth, are there realistic design or statistical approaches that can be employed to minimize the misattribution, including more cohorts of students, a classroom variance component, or including a classroom covariate?

Methods

To answer these research questions, I used two types of simulated data to demonstrate that classroom disturbances result in misattribution of classroom characteristics to the teacher effect in these four VAMs; examined the bias in the teacher effect from the classroom disturbance; compared the performance of these models in estimating a simulated and known true teacher effect; and proposed and tested three design and method-based solutions to the problem that have the potential to minimize bias. In testing these solutions, I maintained the following practical constraints on the evaluation system: first, the number of years used to estimate a TVA score had to be low enough to be timely, particularly for less experienced teachers or teachers new to North Carolina who would have few years' data available; second, that the models be estimable on a system that a state could realistically be expected to have access to; and third, that the data consisted of those typically included in administrative data systems (*i.e.*, not surveys).

Value-Added Models

Random Effects Model (HLM3). A random effects model treats the multiple levels of data—student, classroom, teacher and school—as hierarchically nested units, where each lower level (*e.g.*, student) can only be nested within one higher level unit (*e.g.*, classroom). These include hierarchical linear models and multilevel models (Raudenbush & Bryk, 2002). In this study, I estimated the following multilevel model:

$$(5) \qquad Y_{itsw} = X_{its}\beta_{ts} + X_s\beta_s + \beta_{w-1}Y_{i,w-1} + \beta_{w-2}Y_{i,w-2} + u_s + u_{ts} + e_{its}$$

Subscripts indicate the student (*i*), teacher (*t*), school (*s*) and period (*w*). The variable Y is a student test score on a standardized end-of-grade exam; X_{its} is a vector of student characteristics that are associated with the accumulation of knowledge, as well as a constant; X_s is a vector of school characteristics; w = the period for which the teacher effect is being estimated; *w-1* = the prior grade level; *w-2* = the grade level two years' prior. Therefore, Y appears both as the dependent variable in the current period, as well as predictors of student achievement in the current period as both one-year and two-year lags or pretests. The terms u_s, u_{ts}, and e_{its} are errors for (in order) the school, the teacher, and the student. The teacher effect was estimated using an Empirical Bayes estimator as follows, with b_{0j} a measure of the intercept vector of β_{ts}:

$$(6) \qquad b_{0j} = \gamma_{00} + u_{ts}$$

However, student scores were centered at the grand mean, γ_{00} = 0, implying that b_{0j}= 0 + u_{ts}= u_{ts}. Therefore, the teacher effect was estimated using the Empirical Bayes' residual estimate of u_{ts}.

A variety of random coefficients models have been tested in the literature including by Henry, Rose & Lauen (under review), Guarino et al. (2010) and McCaffrey et al. (2004). Among several random effects models tested by Henry, Rose & Lauen, the model specified in equation 2 was the best performer among simpler variations that excluded the school random effect or the two-year lag.

To satisfy the unconfoundedness requirement of the RCM, the u_{ts}, the estimates of which comprise the teacher effects, had to be conditionally random (*i.e.*, not associated with any other terms predicting student achievement). I assumed that the confoundedness of non-

random assignment was controlled for via inclusion of pretests and covariates that predict student learning, such as family income. This assumption may or may not be valid (*e.g.*, Chetty, Friedman & Rockoff, 2012; Rothstein, 2010). To satisfy SUTVA, the influences not attributable solely to the teacher, such as those from interactions, must not exist or would have to be attributed to the student or to the school, requiring that all system-wide student-teacher interactions be homogeneous, which is unlikely.

Student fixed effects model (SFE). The second major type of VAM examined was a fixed effects model. Fixed effects models are econometric models in which students or teachers (or both) are used, in the context of panel data comprising multiple years (*e.g.*, three), as their own controls. Fixed effects models can be estimated using dummy variables for each "unit," which in this case is a student or a teacher; or they can be estimated by differencing methods, which are similar to mean-centering methods except that both predictors and outcomes are mean-centered (Allison, 2009).

In this study, a student fixed effect model (SFE) was used to estimate a teacher effect from the value-added specification in equation 1. Achievement and all of its correlates were de-meaned (i.e., mean-centered) that is, the mean of each measure was subtracted from the value of each measure in each of the three years of the panel. These demeaned terms were entered into a regression model as follows:

$$(6.1) \quad (Y_{it} - \overline{Y}_i) = (\mu_{it} - \overline{\mu}_i) + (\alpha_i - \overline{\alpha}_i) + (e_i - \overline{e}_i)$$

The term μ_{ti} represents time-varying or contemporaneous predictors of student achievement. The student fixed effect α_i, comprising characteristics of the students that did not change during the multiple years, was accordingly eliminated by de-meaning ($\alpha_i - \overline{\alpha}_i = 0$). This elimination satisfied unconfoundedness for these invariant characteristics, but had

no bearing on the influence of time-varying or contemporaneous characteristics such as the classroom effects. Model 6.1 reduced to model 6.2 as follows:

$$(6.2) \quad Y^* = \mu^* + 0 + e^*$$

The teacher effect was estimated as the mean of the composite student residuals, $(e_i - \bar{e}_i) = e^*$, within each teacher. To satisfy the unconfoundedness requirement of the RCM, this complex error had to be unassociated with the time-varying inputs that were not eliminated by de-meaning. Satisfying SUTVA required the same assumptions as in the HLM3.

Hybrid fixed and random (URM). The Univariate Response Model (URM) is a variation on the simple random effects model that incorporates a function of three to four previous achievement scores in the model to estimate the teacher effect (Wright, White, Sanders & Rivers, 2010). The process by which the composite C_i of prior achievement is estimated is a multi-step process, and summarized in Appendix 1. The process involves multiple phases of de-meaning, and therefore makes the model similar to a fixed effect model. However, it is also similar to a random effects model in that teacher effects are estimated using Empirical Bayes' residuals. The model is as follows:

$$(10)\ y_{it} = \beta_0 + \beta_1 C_i + u_t + e_{it}$$

The nesting in this final model was of students within teachers within one school year. Covariates besides the pretests were not included under the assumption that multiple pretests spanning two years and two subjects control for all confoundedness (Ballou, Sanders & Wright, 2004). This assumption may not be valid (*e.g.*, Amrein-Beardsley, 2008). The assumptions of the RCM will be satisfied only if this is true. Satisfying SUTVA required the same assumptions as the HLM3 and SFE.

Dynamic ordinary least squares (DOLS). A pooled ordinary least squares regression model as described in Guarino, Reckase & Wooldridge (2012) was tested as follows:

$$(11) \qquad Y_{it} = X_{it}\beta_t + \beta Y_{i,w=1} + e_{it}$$

Unlike typical models that are estimated on panel data of multiple years, such as fixed effects or repeated measures random effects models, the panel nature of the data was ignored and each observation on each student, which was not actually independent from other observations on the same student, was treated as independent. The method of estimating the teacher effect was similar to the SFE; the errors for students within each teacher were averaged.

To satisfy the unconfoundedness assumption of the RCM, the error must not be associated with any of the other terms in the model. To satisfy SUTVA, the same assumptions as required by the HLM3, SFE and URM are required. Despite the simplicity of the DOLS, in a comparison of the DOLS with several fixed and random effects models, Guarino, Reckase & Wooldridge (2012) showed that the DOLS was the best performing under a variety of student-teacher assignment mechanisms.

Actual North Carolina Data

Several stages of this simulation study required actual student and teacher data. These stages included a calibration study for setting the inputs in the data generation process in the simulations, as well as an analysis to examine whether some of the models proposed could be realistically estimated. I used a statewide data set of students, teachers and schools in North Carolina public schools during the years 2006 to 2010. These data were provided by the Carolina Institute for Public Policy (CIPP), which obtained the raw data files from the North

Carolina Department of Public Instruction. The data were limited to fifth grade students in the years 2008 to 2010, as these students correspond to the population being simulated. Several student, classroom, teacher and school covariates were available in these data; these covariates are listed in Table 1.2.

Data Generation Process

Data were simulated to provide a "true" teacher effect benchmark against which to compare the estimates of each VAM. Two simulations were conducted, one to test a violation of each of the two central assumptions of the RCM examined in this study. First, a simulation based on variance decomposition was conducted to test the violation of SUTVA in the absence of confounding effects. Second, a simulation based on correlated covariates was subsequently estimated to test the violation of unconfoundedness.

Although every effort was taken to generate data that were as realistic as possible, several simplifications were employed that made the data generation process and estimation more tractable, and the results more easily interpreted. First, the value-added component to student learning, or the persistence of previous years' teacher effects, was left out of the data generation process. Without this simplification, classroom effects in previous periods would propagate into later periods, confounding my effort to estimate the influence of the contemporaneous disturbance on the TVA. A robustness analysis, which I describe in the limitations section, re-examined this decision. Second, for computational feasibility, the number of students, teachers, and schools generated was smaller than that of a typical state. Third, school and LEA membership of both teachers and students were fixed and consequently neither students or teachers could change school or LEA. Fourth, only one grade level was used to assess teacher effectiveness (fifth grade). Fifth, missing data were not

simulated. Sixth, only one subject was used at a time as an outcome, though depending on the calibrated values of the inputs, the results could be interpreted as typical for either math or reading outcomes. The third through sixth simplifying procedures rendered the simulations less complicated.

Data generation by variance decomposition. A simulation without fixed effects and with random assignment of students to teachers allowed me to examine the influence of SUTVA as a classroom disturbance without the additional influence of confounded non-random assignment on the classroom disturbance. Each level of the data generation process—student, classroom, teacher, school and LEA—was assigned a pre-specified variance component φ_{mk}^2, with m = {i = student, c = classroom, j = teacher, s = school, d = LEA}, k = subject = {math, reading}, and w = time or grade level. The classroom effect had a w subscript, while all others were assumed time invariant. Each variance component represented that level's proportion of overall variance in the outcome. None of these generated variables were correlated with each other. These variance components were then square rooted to obtain standard deviations φ_{mkw} and multiplied by standard normal random variables ω_{mk} that introduce variability between simulations, as follows (leaving out subscripts):

$$(12) \qquad \phi = \omega\varphi$$

Each of these components was a standard deviation that was assumed to be the effect associated with that level, or in other words, the contribution of that particular level—student, classroom, teacher, school or LEA—to the overall test score for any student. The true teacher effect was thus the resulting standard deviation for the teacher. The mean for each of three standard normal test scores, one for each grade level, was specified and then

added to the student, classroom, teacher, school, and LEA effects created via the variance decomposition, as well as a random or measurement error component r_{icjsdkw}, to arrive at the total score for each student in each grade level, as follows:

$$(13) \quad Y_{\text{icjsdkw}} = \mu_{jw} + \phi_{ik} + \phi_{ckw} + \phi_{jk} + \phi_{sk} + \phi_{dk} + r_{\text{icjsdkw}}$$

The teachers were assumed to teach one group of 17-23 students in any given year, and peer groups were randomly re-sorted from year to year. The original purpose of this simulation was to provide information to the NC Department of Public Instruction, and the number of students and teachers was selected to reflect a range of actual districts of different sizes. Thus I simulated 833 teachers and 16,542 students in 827 classrooms, with 17-23 classrooms per teacher. I started with two cohorts of students. The data were entered in the form of z-scores with a mean standardized test score in each year of 0, and a total standard deviation of 1, absorbing all sources of variation as specified in model 13. In the baseline scenario, classroom variance was held to zero.

Data generation by correlated covariates. A simulation having correlated covariates as characteristics of students, classrooms, teachers, and schools was devised such that the confoundedness of non-random assignment could be examined as the source of the classroom disturbance on TVA estimation. A correlation matrix decomposition procedure (Vale & Maurelli, 1983) allowed me to specify the exact correlations between simulated student, classroom, teacher and school fixed effects. The resulting classroom and teacher effects varied within their respective contexts; *i.e.*, the simulated effect for any given classroom varied within that classroom. To solve this problem, I averaged the effects up to their respective levels. This process slightly altered the correlations but not enough to substantively alter the findings.

Each level of the data generation process was assigned one or more correlated covariates μ_{mkw}. Correlations between student and school, student and teacher, classroom and school, and teacher and school were set at .20. The correlation between students' teachers over time was set at 0.50. These were realistic values as determined by a calibration study (see below). The correlations between the classroom effect and student effect, and between the classroom effect and teacher effect, were varied to examine the bias in the TVA due to classroom disturbances of different intensities; this varying intensity is discussed below. Level-specific standard deviations were also generated in a manner similar to that described above for the variance decomposition simulation. These included η_{ik} for student, η_{ckw} for classroom, η_k for teacher, and η_{sk} for school. A subject and grade-specific state grand mean (μ_{kw}) and residual ($e_{icgsdkw}$) were also estimated. The resulting fixed effects and standard deviations were all added together to produce the student achievement level in a given year, as follows:

(14) $$Y_{icgsdkw} = \mu_{kw} + \mu_{ik} + \mu_{ckw} + \mu_{jk} + \mu_s + \eta_{ik} + \eta_{ckw} + \eta_{jk} + \eta_{sk} + e_{icgsdkw}$$

In this form of the data generation process, the variance component was included as parameter error, rather than as the effect for that level. This decision was re-examined in a robustness analysis. The "true" teacher effect was the simulated fixed effect for each teacher. In this design, students remained with their peer groups as they advanced. For the sake of simplicity, I simulated 1,000 teachers and 40,000 students in 2,000 classrooms, with 2 classrooms per teacher (2 cohorts). Z-scores were used in this simulation as well. In the first set of models, the classroom covariate had a zero correlation with the other fixed effects, but the other correlations were non-zero.

Calibration of inputs. To develop simulations that were as realistic as possible, I used two sources of information to determine the levels of realistic inputs, including those for the teacher effect and its relationship with the classroom effects, for the two data generation processes. First, for the variance decomposition simulation, I required a proportion of variance at the student, teacher, and school levels. A proportion of variance at the classroom level was not needed as my research question implies that the classroom variance may absorb all of the teacher variance. I examined actual fifth grade NC data on math and reading achievement. Depending on whether the models were conditional or unconditional or spanned one or multiple years, between 10 and 25% of the variance in student outcomes was at the teacher level (Table 1.3). Most of the remaining variance was at the student level (from 0.53 to 0.89), with the balance carried by the school level (0.06 to 0.20). According to a Nye, Konstantopolous & Hedges (2004) study of variance decomposition, teacher variance should be around 11%, though the grade levels examined were lower (first through third) than intended here (fifth). I set teacher variance to be 14% of total variance in the outcome. The student and school proportions of variance were set to 0.77 and 0.08; classroom variance was zero until introduced as a disturbance.

Correlated covariates were informed by additional analyses on actual North Carolina data. The key inputs to the correlated covariates model were the levels of the correlation between student, classroom, teacher, and school covariates in the data generation process. Using the fifth grade data on math and reading achievement as well as student, classroom, teacher and school covariates, I observed that typical correlations between student covariates and all other levels was as high as 0.55; between classroom and all other levels were as high as 0.20; and between school and all other levels was as high as 0.55. The parameter variances

(the amount of conditional variance at each level) were informed by the calibration for the variance decomposition. For students, parameter variance was set to 0.77; at the classroom level to 0.02; at the teacher level to 0.12 and at the school level to 0.08. The correlation of each student's third grade teacher with their fourth and fifth grade teachers was set sufficiently high (0.50) such that students had a high probability of being assigned to a teacher of similar skill level in each year.

Number of simulations. The number of simulations in this study, 100, was low relative to what is recommended in general for simulations (normally in the thousands) but tradeoffs were necessary due to the high computational demands of some of the models and the length of time it took to finish a complete round of models. As the typical use of large N in a Monte Carlo simulation is to smooth out the variability between simulations imposed by measurement or random error, I elected to minimize the random error simply by specifying that it should only constitute 1% of the variance in the student outcome. A robustness check, conducted on only one classroom variance scenario comparing the results of 100 simulations with that from 1,000 simulations, demonstrated that the two were nearly identical.

Classroom disturbance input. Using the simulated data, I examined and compared VAMs for estimating teacher effectiveness under a range of scenarios having different classroom effects disturbances. First, in the variance decomposition simulation, I created seven scenarios in which I varied the proportion of variance attributable to the classroom from 0 to 0.12 in 0.02 increments. The rationale behind varying the classroom disturbance in this range was to test the effect on the TVA estimate if the teacher effect could be almost entirely explained by interactions in the classroom. The teacher variance was accordingly reduced by the same two units, starting at 0.14 and varied down to 0.02. To have a teacher

effect to estimate, the lowest teacher proportion of variance was set 0.02. In effect, the two components were swapped in 0.02 increments. Total variance was not changed and remained at 1 throughout.

Second, in the correlated covariates simulation, I applied varying intensity to the correlation between the classroom effect and student and teacher effects, creating scenarios in which the correlations had the following values: -0.6, -0.4, -0.2, 0, 0.2, 0.4 and 0.6. This range covered the range of observed correlations between student, classroom, teacher and school covariates as observed in the calibration study.

Methods to Minimize Classroom Effects Disturbances

After comparing the four value added models, I propose and examine the three methods, proposed above, which may reduce the misattribution of classroom variances and effects: a) adding additional cohorts of students to the VAM estimation (up to 5); b) testing a four level model of students, classrooms, teachers and schools (a fifth VAM, with the classroom level as the enhancement); and c) testing models with a classroom covariate added. In all these cases, the extent of disturbance in biasing the teacher effect was examined, and the performance of each model is compared with the true effect.

Additional cohorts. Two cohorts of students were used in the baseline simulations. In an elementary school setting where a classroom of students is typically with the same teacher throughout the course of the day, a cohort of students is equivalent to a year observing the teacher in one classroom setting. As cohorts are increased, the number of years and thus the number of classroom settings is increased. The increased number of classrooms represents a larger sample of teacher-classroom measurements for each teacher, and large sample theory suggests that as long as there is no systematic bias over time, the average of

the TVAs over the multiple cohorts should converge on the true value. Although it would be ideal from a measurement standpoint to allow the number of years to be large (*e.g.*, 20), from the perspective of informing teacher evaluation, which requires timely data on all teachers including younger, less experienced teachers, the number of years should be limited. Henry, Bastian & Fortner (2011) demonstrated that teachers' effectiveness grows over the first three to five years. In this analysis, I doubled the number of cohorts from two to four.

Four-level model (HLM4). A four-level model was estimated, adding a classroom variance component u_{cts} to the HLM3 as follows:

(15) $$Y_{itsw} = X_{its}\beta_{ts} + X_s\beta_s + \beta_{w-1}Y_{i,w-1} + \beta_{w-2}Y_{i,w-2} + u_s + u_{ts} + u_{cts} + e_{its}$$

In the HLM3, just as in the SFE, URM and DOLS the classroom disturbance did not have a parameter of its own to manifest on and therefore manifests on another parameter, biasing that parameter as an estimate of its true value. Alternatively, in the HLM4 specified in Model 15, the between-student but within-teacher variance that contains the classroom disturbance is attributed to a separate variance component u_{cts}. Because a four level model is more computationally intensive than a three level model, there was some concern that estimating such a model would be impractical. I therefore used the CIPP actual NC data to test a simple statewide HLM4 over a period of 3 years to determine if the computational intensity of the model presented a practical constraint on implementation of such a model.

Classroom covariate (HLMC). As an alternative to the classroom variance component, which may be too computationally demanding to be implemented in a typical statewide teacher evaluation system, I proposed to include a classroom covariate in the HLM3. Because the variance decomposition simulation has no fixed effects because students are randomly assigned in that simulation, this strategy was only available for the correlated

covariates simulation. Classroom covariates are often simple aggregates of student characteristics (*e.g.*, Zimmer and Toma, 2000). The classroom covariate simulated in the correlated covariates data generation, which had a correlation with the student and teacher effects of varying intensity, reflecting the calibration study, of -0.6, -0.4, -0.2, 0, 0.2, 0.4 and 0.6, was entered into the models. As noted, there are two opportunities in this solution for biasing the teacher effect. The covariate may not sufficiently control for classroom effects, and therefore the teacher effect may still be biased after inclusion of the covariate. Further, the classroom covariate may also have shared variance with the teacher effect. If the classroom covariate is modeled, this shared variance will be removed from the variance used to estimate the teacher effect, potentially biasing the teacher effect.

The covariate, the μ_{ckw} generated in the simulation, was added to the HLM3 as follows, with coefficient β_{μ}, to yield the HLMC model:

$$(16) \quad Y_{itsw} = X_{its}\beta_{ts} + X_s\beta_s + \beta_{w-1}Y_{i,w-1} + \beta_{w-2}Y_{i,w-2} + \beta_{\mu}\mu_{ckw} + u_s + u_{ts} + e_{its}$$

Comparison Methods and Criteria

Mean squared error of TVA. To assess the impact of unaccounted-for classroom variance and correlated covariates on the estimation of teacher effectiveness, the mean squared error was calculated. The MSE is the average of the squared differences between the TVA and the true teacher effect:

$$(17) \quad MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{T} - T)^2$$

Criteria for comparison. Three criteria were used to compare the absolute performance of each VAM on estimating the true teacher effects in the simulated data. First, Spearman rank order correlation coefficients, a non-parametric measure capturing the association between the rankings of two variables, was estimated for each pairing of a VAM

with the true effect and with each other VAM. This simple measure describes the relative placement of teachers along the entire distribution of teacher effectiveness as determined by the true and TVA estimates, regardless of the absolute differences between teachers' placement in these two distributions. For the simulated data, the estimates in each simulation needed to be combined into a single point estimate which required a Fisher z transformation; the mean of this z-transformed correlation was calculated, and then back-transformed using the *tanh* function (Schafer & Graham, 2002). High performing VAMs have relatively higher Spearman coefficients.

Second, I calculated the percent agreement on fifth percentile teachers: First, the teachers in the bottom 5% under the "true" effect were identified, as were the teachers in the bottom 5% on each TVA. Subsequently, those who were similarly classified (in the bottom 5 or above the bottom 5) under both versions were considered "in agreement" while those who were in the bottom 5 on one but not both were considered not in agreement. A 5% threshold is important because this represents a relatively small proportion of teachers who are likely to be considered for the harshest sanctions or most demanding coaching and assistance. High performing VAMs have relatively higher levels of agreement. Because the simulated distributions were normal, the findings were nearly identical for teachers in the 95th percentile.

Third, I examined the false identification of ineffective teachers in the simulated data. First, I identify those teachers above the cutoff for ineffectiveness (using a z-score of -1.64 or -1.64 standard deviations from the mean, implying that 95% of teachers should be above the cutoff) on the "true" measure. Then I identify the proportion of teachers who were below this cutoff on the estimated TE. The teachers who satisfy both conditions were considered false

positives or falsely identified as ineffective. This method of comparison differed from the percent agreement method, despite having a very similar threshold for ineffectiveness, because the focus was on the experiences of average teachers found ineffective, rather than on the overall level of agreement. High performing VAMs have relatively low proportions of false positives. In addition to the proportion, I calculate how many of North Carolina's approximately 9,000 fifth grade teachers would be falsely identified as ineffective by each TVA, and I calculated the mean true scores of teachers falsely identified as ineffective. Note that for the opposite problem—falsely identifying a teacher as highly effective when he or she is not, which although certainly important is not in my view as risky as falsely identifying ineffective teachers—I could assume that the results would actually be very similar as the simulated distributions are highly normal. Note also that given the distributional rather than absolute threshold of effectiveness, the number and proportion of teachers falsely identified as ineffective imply an equal number and proportion of teachers who were ineffective but identified as average or better, as well as an equal number of teachers falsely identified as average or above average. Thus, the true disturbance on the teacher workforce could be four times as high as that reported in the tables.

Results

Variance Decomposition & SUTVA

MSE of the teacher effect. The MSE of each TVA increased approximately linearly as the classroom portion of variance was increased (Table 1.4 and Figure 1.1). At zero, the HLM3 was the best performing model (MSE = .011), followed closely by the URM (MSE = 0.014). The SFE (MSE = 0.021) and DOLS (MSE = .026) performed relatively poorly. However, as the proportion of classroom variance increased, the SFE, though it increased,

did not increase as much as the other three models, and when the proportion of classroom variance reached 0.06, the SFE had the lowest overall MSE. The relative ordering of the other three models was preserved.

Spearman rank order. When the classroom variance was held to zero, correlations of each of the four models with the true effect was very high at .955 for the HLM3; .941 for the SFE; .946 for the URM; and .909 for the DOLS (Table 1.5). As the classroom effect disturbance was increased from zero to 0.12, these correlations decline. The trend in these correlations as the classroom proportion was increased took on a convex curvilinear shape (Figure 1.2), with the degradation in the TVAs increasing relative to equivalent increments of the classroom effect. At 0.12 of variance (at which point only 2% can be attributed to the teacher), the TVAs have much more modest rank correlations with the true effect, with the HLM at .460; SFE at .454; URM at .456; and DOLS at .437.

Percent agreement on fifth percentile. When classroom contributed zero variance, agreement on classification in the bottom 5% was very high, with the HLM at 97.71%; SFE at 97.33%; the URM at 97.44%; and the DOLS at 96.62% (Table 1.6). As classroom variance increased (Figure 1.3), performance declines nearly linearly for all four models, with a drop of just under one percentage point in agreement for each two percentage point increase in the proportion of variance due by the classroom. The differences between the models were modest (no more than 0.8 percentage points).

False identification as ineffective. When the classroom proportion of variance was set to zero, 2% or less of average or better teachers were falsely identified as being ineffective at a threshold of -1.64 standard deviations (numbers of teachers out of 9,000 in parentheses): for the HLM, 1.6% (n = 144); for the SFE, 1.8% (n = 160); for the URM, 1.7%

(n = 155); for the DOLS, 2.0% (n = 180) (Table 1.7). As the proportion of variance attributed to the classroom increased to 0.12, the four models converged to 3.9-4.0% falsely identified, or about 352-358 out of 9,000 teachers. The true scores of teachers falsely identified as ineffective increases with a slightly convex shape indicating greater degradation in performance as the proportion of the variance attributed to the classroom increased in equivalent increments (Table 1.8). At a classroom variance of zero, the mean true z scores were -1.43 for the HLM3; -1.40 for the SFE; -1.41 for the URM; and -1.33 for the DOLS. Higher scores implicate the TVA in more egregious errors, as teachers farther from and above the cutoff point for ineffectiveness are wrongly found to be ineffective. At 0.12 of variance at the classroom level, the four models converged to a range of -0.63 to -0.58 mean true scores (Figure 1.4).

Correlated Covariates and Unconfoundedness

MSE of the teacher effect. The mean squared error in the teacher effect remains roughly constant as the correlation between the teacher and classroom effect was varied from -0.60 to 0.60 (Table 1.9 and Figure 1.5). The DOLS had the highest MSE, between 0.13 and 0.14; the HLM, SFE and URM were all roughly similar, between 0.05 and 0.10, and varying only slightly.

Spearman rank order. When correlation between classroom and teacher was held to zero (though teacher-student, teacher-school, student-school and classroom-school correlations were non-zero), the HLM was the best performing model at a 0.788 (Table 1.10), with the DOLS (.701), URM (.701) and SFE (.639) following. Varying the correlation between classroom and teacher from -0.60 to 0.60 did not greatly alter these findings, with

each model changing only slightly as the intensity of the correlation moved away from zero in either direction (Figure 1.6).

Percent agreement on fifth percentile. At zero correlation between classrooms and teachers, the four models had similar performance in finding teachers in the bottom 5% of the distribution with the HLM at 94.8% agreement; SFE at 93.4%; URM at 93.9%; and DOLS at 93.9% (Table 1.11). Further, performance did not change very much as the correlation was varied between -0.60 and 0.60 across all of the models (Figure 1.7).

False identification as ineffective. At zero correlation between classrooms and teachers, 2.6% of teachers would be misidentified as ineffective (Table 1.12) using the HLM (n = 232 out of 9,000 teachers); 3.3% using the SFE (n = 293); 3.1% using the URM (n = 277); and 3.1% using the DOLS (n = 275). There were no substantive changes as correlation was increased in either direction to 0.60 or -0.60. The true scores of teachers misidentified as ineffective ranged between -1.12 (HLM3) and -0.96 (URM) with the DOLS (-0.99) and SFE (-0.89) in between. These were also largely invariant to the level of correlation between classroom and teacher effects, though true scores tended to decline as the correlation moved from -0.60 to 0.60, and tendency that was most substantial for the SFE (Table 1.13 and Figure 1.8).

Proposed Solutions

Additional cohorts of students. When the number of cohorts in the variance decomposition simulation was doubled from two to four, neither the Spearman rank order or percent agreement on fifth percentile were any different from the results obtained using two cohorts (Tables 1.14 and 1.15, respectively; Figures 1.9 and 1.10, respectively). However, the proportion and number of teachers falsely identified as ineffective was less than half of that

when using two cohorts (Table 1.16). When classroom variance comprised 0.04 of overall variance, the models showed that under the HLM, 0.9% (n = 85) of teachers would have been misidentified as ineffective; under the SFE and URM, 1.0% (n = 89 and 88, respectively, the differences due to rounding error in the proportion); and 1.1% (n = 98) under the DOLS. When the proportion of classroom variance was increased to 0.12, the proportion and number remained below half and the average true score moved farther away from the cutoff (Table 1.17 and Figure 1.11).

When the number of cohorts in the correlated covariates simulation was doubled from two to four, and when the correlation between classrooms and teachers was moderate and well within the observed range (0.40), there was small improvement in most models. On the Spearman rank correlation, the HLM3 improved from .762 to .820; the SFE improved from .638 to .706; however, the URM fell from 0.686 to 0.748. The DOLS went from 0.697 to .748 (Table 1.18). On percent agreement on the fifth percentile (Table 1.18), there was negligible improvements in all models; the HLM3 went from 94.68% to 95.32%; the SFE from 93.49% to 94.07%; the URM from 93.87% to 94.50%; and the DOLS from 94.05% to 94.45%. On the proportion and number of teachers falsely identified as ineffective (Table 1.18), the proportion and number of teachers changed only slightly, with the HLM3 improving from 2.7% (n = 245 teachers) to 2.4% (n = 212); the SFE falling from 3.3% (n = 299) to 2.9% (n = 264); the URM falling from 3.2% (n = 284) under 2 cohorts to 2.8% (n = 249) under 4; and the DOLS falling from 3.0% (n = 270) to 2.8 (n = 252). The average true scores, similarly changed only very slightly.

Four level multilevel model (HLM4). A four level HLM was added to the models to determine if a separate variance component for the classroom level would minimize the bias

in the TVA. In the variance decomposition simulation, the HLM4 had the same or only slightly higher (*e.g.*, 0.01 difference) Spearman rank order correlation than the HLM3, generally the best performing model (Table 1.19). For example, when classroom variance was 0.04 of the total, both the HLM3 and HLM4 had a Spearman rank order coefficient of 0.864. Alternatively, when classroom comprised 0.12 of total variance, the HLM3 was at 0.460 and the HLM4 at .461. A similar finding emerged for percent agreement on the fifth percentile (Table 1.19). On the false identification of ineffective teachers (Table 1.19), at both 0.04 and 0.12 of variance, the HLM3 and HLM4 were roughly equal in proportion and number misidentified, as well as the average true score. No changes were observed for the correlated covariates simulation when the HLM4 was introduced (Table 1.20).

The modest-to-nonexistent improvement in performance of the HLM4 over the HLM3 may not be worth the computational requirements of the HLM4. An HLM4 estimated on actual as opposed to simulation data might be too computationally intensive to estimate on a system likely to be available for a state teacher evaluation system. To ascertain whether a four level model could be estimated on such a system using real data, I used actual NC data to estimate a conditional four level model with student, classroom, and school covariates. I used three cohorts of students to estimate a fifth grade teacher effect. The operation could not be completed with available memory. I attempted the operation again with one cohort at a time, but this too could not be completed. This suggests that even if the classroom variance approach had worked to adequately minimize the disturbance in the TVA, such a strategy may not be possible in practice, at least at the present time.

Classroom covariate (HLMC). In the correlated covariates models, I added the classroom effect μ_{ckw} to the HLM3, the best performing model in this analysis. The results

depended on the level of correlation between the classroom and teacher effects. When the correlation was zero, the Spearman rank correlation changed negligibly. Although the HLMC was a better performing model than the SFE, URM and DOLS in every scenario, it did not improve the resilience of the random effects framework to estimate an unbiased TVA. A similar pattern emerged for the percent agreement on fifth percentile, with the HLMC model performing slightly better than the HLM3 for scenarios with low correlation in between classroom and teacher, and slightly worse at high correlations (Table 1.20).

On the false identification of ineffective teachers, the results were similar, with the HLMC (2.4%; n = 220) being only slightly better than the HLM3 (2.5%; n = 223) at zero correlation, but then slightly worse at all other correlations (Table 1.20). The difference was, however, negligible.

Discussion

Value-added models are intended as a means of estimating teacher effectiveness using changes in student achievement on end-of-grade standardized tests. These models use econometric or statistical methods to adjust for inputs to learning that are not directly related to the teacher in an effort to derive an unbiased measure of teacher effectiveness. The use of outputs rather than inputs in this manner addresses research that shows that typical inputs to teacher training are not highly associated with student learning (*e.g.*, Boyd et al., 2009; Henry et al., 2011). Coupled with the availability of standardized test data and federal and state mandates for reform, states are implementing statewide teacher evaluation systems that include the use of VAMs.

This study examined four VAMs and provided evidence that all of them estimate teacher effectiveness with non-zero error; that one model in particular, the HLM3,

outperformed the others regularly; and that the resilience of these models to scenarios of increasing classroom disturbances depended both on the model and the type of disturbance. For the most part, the performance of these models did not vary in accordance with varying correlation between the teacher and classroom effect in correlated covariate simulations. The evidence from the zero-correlation scenarios in contrast with the scenarios in the variance decomposition simulation suggests that these TVAs are greatly affected by the presence of correlation and shared variance between teachers and their students and schools. The classroom disturbances did not contribute further bias. However, in the simulations for variance decomposition for assessing the impact of SUTVA violations, all of the models experienced substantial degradation.

Although the Spearman rank order correlation provided reliable evidence that the HLM3 was the top performer, under certain classroom effect scenarios, it did not differ enough to favor this model over the others when the full range of classroom effects scenarios was considered. Further, on the MSE, the SFE was a far better model as the classroom effect increased. When observing percent agreement on the bottom 5% of teachers, the results were not too different from those observed on the Spearman rank order. In this comparison, the models were bunched together tightly, with the HLM3 slightly outperforming the others. However, the models converged to a point as the classroom effect was increased. The false identification of ineffective teachers showed that these models performed similarly, though again slightly favoring the HLM3. Of the approaches that I tested to reduce the effect of these disturbances, only increasing the number of cohorts showed potential in reducing the bias of the TVA and the false identification of ineffective teachers. However, the benefits of this approach were nevertheless modest and subjected to a strong caveat. A substantial

amount of bias and misidentification remained, and opportunities for further reductions are limited by the need to provide timely information on new and inexperienced teachers. Given the number of years of data needed to obtain relatively reliable TVA estimates, the evaluation of teachers who are new to their subject areas may be challenging. It is also worth noting that in the course of conducting multiple versions of these analyses, often the results changed by a greater amount as a result of simply random variation, than they did as a result of the solutions that I proposed.

Limitations

The key limitation in this study is that I utilized stylized simulation data to examine the influence of a single contemporaneous disturbance, the classroom effect in the current period, on the estimation of TVAs. This required generation of data that isolated this disturbance in the current period free of other pathways for this disturbance to further influence TVA estimation. An important component of the process of student learning, that new knowledge generated in each period is added to that gained in the previous period, was left out of the data generation process. This component, a central part of the learning process as expressed in the education production function, is the value-added that gives value-added models their name, and thus this approach may seem novel and potentially counterintuitive to satisfying the objectives of a study of VAMs. However, if these effects had been included, then the classroom effects of previous periods would have been propagated through the value-added component. Leaving this component out allowed me to isolate the influence of this effect on measurement of TVAs absent the influence of past years' classroom disturbances, but it did so at a cost of sacrificing some realism to the data. Under more real

conditions, the performance of the models in estimating TVAs may be degraded further, and these degradations may not be preserve the rank ordering of the four models.

Robustness checks showed that for some of the models, this distinction did not matter. In a model with classroom and teacher effects correlated at 0.20, the HLM3 and URM, demonstrated the same level of performance in estimating the true effect as measured by the Spearman rank order correlation as when the value-added component was excluded. The performance of the SFE and DOLS, on the other hand, was further degraded by the inclusion of these learning effects. The Spearman rank order correlation for the SFE fell from 0.64 to 0.56, and for the DOLS it fell from 0.70 to 0.49. The primary difference between the two groups of models was that in the HLM3 and URM the variance was decomposed into constituent parts first. In the case of the HLM3, variance was decomposed into student, teacher and school. In the case of the URM, school effects on prior achievement were controlled for via fixed effects, and the teacher effects were subsequently estimated from teacher-level variance separate from student-level variance. On the other hand, in the SFE and DOLS, teacher effects were estimated by averaging up student-level residuals to the level of the teacher. If learning effects manifested on the error, given non-randomness in the distribution of these effects across classrooms, the aggregating process may have attributed prior learning to the teachers. This would have occurred even in the presence of an explicit control for previous achievement as in the DOLS, if there was a difference between within-teacher (controlled for) and between-teacher (not controlled for) effects.

The inclusion or exclusion of the learning effect is not simply a technical question but may have some substantive support. The exclusion of the previous period's learning effect is equivalent to a zero-persistence model of student learning in which all knowledge

accumulated in the previous period is lost. Although this seems unrealistic, one study demonstrated that the persistence of teacher effects is very low, with 80% of student learning gains in one period decaying into the next (Jacob, Lefgren & Sims, 2010). Studies of VAMs generally use the education production function, which depicts learning as a cumulative and even additive process, as a basis for assuming some form of persistence in learning from period to period (*e.g.*, Guarino, Reckase & Wooldridge, 2012; Jacob, Lefgren & Sims, 2010). McCaffrey, Lockwood, Koretz, Louis & Hamilton (2004), included persistence as an estimable parameter in the model. The purpose of excluding the value-added in this study was not to lend support to any one view or the other. This study stands out for excluding this question from consideration in the examination of the sensitivity of four models to a contemporaneous classroom disturbance. This exclusion provided greater clarity in answering the question.

Implications for Teacher Evaluation Policy

Because the teacher effect is measured using assessments of students' learning, and because the purpose of the typical teacher evaluation system is to hold teachers accountable for this growth, there is an implied causal relationship between the measure—the student test score—and the key input, the teacher. That is, these evaluation systems are being designed to attribute changes in student achievement to the practices of individual teachers. This intent is an important point. There is nothing inherently causal about the VAMs themselves, but including TVA estimates in the evaluation system implies a causal inference. Failures of these assumptions can then manifest not only as misdirected professional development supports, sanctions or rewards, with effects on student learning, but also as loss of prestige and decreased psychological well-being of teachers (Finnigan & Gross, 2007; Santoro,

2011). Given the causal inference implied by a TVA evaluation system, it is necessary to understand the influences on student learning that are related to these assumptions, how these influences potentially bias TVA estimation, and how teacher evaluation can be implemented to safeguard against these influences.

The RCM provided a framework in this study for understanding what is required of a teacher evaluation system using TVAs (Holland, 1986). In the RCM, each teacher is an individual treatment "administered" to her students, and in any given state there are numerous such treatments. In North Carolina alone, there are more than 9,000 teachers in each grade level. To use the RCM, a counterfactual for each teacher *j*—conditions that would have prevailed in the absence of this teacher—is required. Although this is an abstraction, it nevertheless has practical implications for the development and implementation of an evaluation system if the necessary conditions cannot be supported.

The counterfactual. In a binary treatment condition, the conditions that would have prevailed in the absence of treatment would be labeled the control condition. In a many-valued treatment such as this one, a typical counterfactual would be *.j* (not *j*), and might be implemented using the average teacher effect. However, the counterfactual cannot be identified until the purpose of the evaluation system—the question that the evaluation system is designed to answer—is identified (Raudenbush, 2004). In the context of classroom effects that may occur in conjunction with teachers' effectiveness, there are two possible counterfactual conditions. A between-context counterfactual condition contrasts condition *j* with condition *.j*. A second, within-context counterfactual condition is possible as well. In this counterfactual, the condition *jk*, of teacher *j* in classroom *k*, is contrasted with condition *j.k*, or teacher *j* not in classroom *k*.

The state evaluation systems that are being deployed, such as in North Carolina, are intended to compare teachers statewide. North Carolina's system, in theory, should enable teachers in two far-flung districts in the state to be exchanged for each other with predictable consequences for the achievement of students taught by these teachers. This type of system suggests that the between-context counterfactual is desired, having teacher *j* with counterfactual *.j*. For this to be the case, the teacher effectiveness estimate cannot be biased by conditions specific to the context in which the teacher was evaluated, because that would imply condition *jk*, which does not have counterfactual *.j*. For the between-context counterfactual to be possible, assumptions related to the assignment of students and teachers to each other, and related to the processes of teaching and learning in a classroom, must be made. These assumptions clarify the well-defined states of the counterfactual and therefore have implications for development and implementation of a teacher evaluation system (Morgan & Winship, 2007).

Assumptions. The key requirements of the RCM for supporting the desired counterfactual condition, and therefore supporting the causal inference of TVAs as unbiased estimates of teacher effects are unconfoundedness and SUTVA. These requirements imply assumptions about teacher-student assignment and the processes of teaching and learning, and in the context of these assumptions, the methods and data used to estimate the TVAs. Unconfoundedness requires that student-teacher assignment be effectively random, perhaps conditionally so after adjusting for measured predictors of assignment. Whether conditional unconfoundedness can be supported depends on the measures available in the data for estimating TVAs; how these measures are incorporated into the VAM; whether these measures have sufficient shared variance with unmeasured inputs to learning to control for

the influence of these unmeasured variables; and whether these measures have shared variance with the teacher effect. The evidence in this study suggests that correlations between teachers, students, schools, and classrooms are important and contribute to error in the measurement of teacher effectiveness via TVAs.

The assumption of SUTVA requires that a student's teacher effect be invariant to the different ways in which that student could be grouped with others; if this assumption was not met, then the student would not have one potential outcome for teacher *j*, but as many potential outcomes as there are combinations of peers. The violation of SUTVA is therefore intuitive and the evidence supporting the violation is robust (Reardon & Raudenbush, 2009; Rubin, Stuart & Zanutto, 2004). If peer effects—students' interactions with each other—affect the rate of learning they experience during the school year, or if the other students interact with the teacher in a way that alters the teacher's methods or practices as they affect each student, then SUTVA is violated. The only means by which SUTVA would not be violated in the presence of these interactions is if the effects of these interactions were equivalent across every possible assignment of students, which is no more credible an assumption. The findings from this study are very clear. As the proportion of variance attributed to the classroom rises, the models were degraded, sometimes substantially.

Together, these findings regarding confoundedness and SUTVA reject that the TVA can be clearly interpreted as a between-context measure of teacher effectiveness without caveats. The findings suggest that the TVAs may not actually be valid for the within-context teacher effect estimate either, and that at best the typical TVA estimate might be a hybrid of the within (*jk*) and between (*j*) effects, which does not have a clearly defined counterfactual. Failing to find that these assumptions are realistic raises questions about interpreting the

TVAs as causal estimates of teachers' effects on student learning. Another factor to consider, and one that has not received much attention in studies of TVAs, is whether these assumptions serve the purpose of fairly evaluating teachers.

Fair Teacher Evaluation. Factors that cause the assumptions of the RCM to be violated may nevertheless be important determinants of a teacher's effectiveness. To put it another way, the SUTVA and unconfoundedness assumptions, if they could be satisfied by some design or statistical method, would serve the purpose of *measuring teacher effectiveness* by attributing student performance changes to teachers, but they might not serve the purpose of *fairly evaluating teachers* if in fact some of these factors contribute in deeper ways to teachers' effectiveness.

Education scholars have raised the possibility that the abstract process of teaching and learning represented by the education production function is too simple and unrealistic and have raised questions about the extent to which it excludes from consideration factors that might inform administrators about unique forms of teacher effectiveness or ineffectiveness. The inferential problem presented by VAMs has provided an opportunity to clarify this limitation. Evidence from education research suggests for example, that both cognitive (Hughes, Luo, Kwok & Loyd, 2008; Skinner & Belmont, 1993) and non-cognitive (Hughes & Chen, 2011; Stipek & Miles, 2008) student processes and behaviors inform the unfolding of teacher practices over the course of the school year, implying that with different groups of students, teachers' practices would be different and the resulting level of effectiveness potentially affected as well. Although the practices themselves might be fairly considered part of a within-context teacher effectiveness estimate, and thus not appropriate to consider as a teacher's effect for estimating a between-context TVA, the teacher's skill set that drove

the feedback process, and made this effect possible, should not be discounted. This is a more complex process than that represented by the education production function. More effective teachers may be more skilled at using feedback to inform further instruction; alternatively, less effective teachers may be less skilled at these approaches. In sum, the technical demands imposed by the implied causal relationship between teachers and students may conflict with the complexities inherent in the processes of effective teaching (Kennedy, 2010).

Many of the assumptions needed to interpret TVAs as causal estimates simplify how the process of teaching is represented in measurement, and more complex questions have not been addressed. For example, many students have multiple teachers or a teacher and teaching assistant in the same classroom; others have outside tutoring; still others switch teachers during the year (Valli, Croninger & Walters, 2007). The technical need in these settings involves partitioning learning gains on the subjects taught in these settings to multiple teachers, and an analysis similar to the one in this study may be required to assess the sensitivity of the models to these issues. Alternatively, the issue of fairness concerns whether more effective teachers make better use of these multiple-teacher settings and that perhaps the multiple teachers' effects are emergent.

The measurement versus fairness question extends beyond the issue of multiple teachers to encompass other contemporaneous school inputs. This study demonstrated that separating the teacher's effectiveness from components of student learning that are wholly or partly due to the unique group of students assigned to the teacher for that year is a significant challenge. The day-to-day work of teachers represents numerous opportunities for such effects to emerge. These situations will interfere with or support the practice of teaching, with teachers adapting their lessons to fit the needs or opportunities of the situation

(Kennedy, 2010). Studies of the situational context of teaching are being developed in response to the intense focus of typical teacher effectiveness studies on characteristics of the teacher rather than the characteristics of teaching (Gilbert & Malone, 1995; Hiebert & Morris, 2012; Hill, Rowan & Ball, 2005; Kennedy, 2010). In sum, teacher evaluation systems that rely on TVA estimation may not contain unbiased estimates of teacher effectiveness, and in serving the needs of estimating teacher effectiveness with TVAs they may miss opportunities for a deeper understanding of teacher, or teaching, effectiveness.

Teacher evaluation policy. Teacher evaluation systems that incorporate TVA estimates should acknowledge issues related to the failures of SUTVA and unconfoundedness assumptions to hold, as well as questions regarding whether TVAs can fairly represent teachers' skills. The policy recommendations for teachers that are found ineffective using a TVA estimate have ranged from high-stakes options such as dismissal (*e.g.*, Gordon, Kain & Staiger, 2006) to low-stakes options such as requiring additional professional development and coaching (*e.g.*, Amrein-Beardsley, 2008), Based on the findings of this study that TVAs measure teacher effectiveness with error and a non-zero probability of incorrectly classifying an average or better teacher as ineffective, I propose a number of potential directions for a teacher evaluation system that mitigate misclassification risk. Further, because TVAs provide no information about the skills in which teachers are deficit, I propose a number of different ways in which alternative measures of teacher effectiveness such as observations (Hill, Charalambous & Kraft, 2012; Pianta et al., 2008) and surveys (Koth, Bradshaw & Leaf, 2008; Rowan, Correnti & Miller, 2002) can be incorporated into the evaluation system.

Most importantly, given the non-zero error, high-stakes sanctions such as dismissal should not be based on a TVA estimate without additional data to support the dismissal. For the most part, this recommendation reflects actual policy decisions. The North Carolina teacher evaluation system being deployed in 2012-13 combines a TVA estimate with five criteria from a rubric-based observational instrument. Error in measurement can be mitigated by combining data from TVA estimates with data from unannounced, rubric-based teacher observations or student surveys administered at periodic intervals, and requiring that action be taken only for teachers who are found deficient on a certain number of these measures. There are potentially subjective decisions involved in this approach, such as how much weight to give to each component, but such subjective decisions at the level of the system are fairer to teachers than subjective decisions made by administrators that prevail in the absence of a teacher evaluation system (e.g., Andrejko, 2004).

Alternatively, a number of low-stakes options are available that make best use of the limited information about teacher performance that TVAs provide. It can be argued, for example, that using TVAs to identify teachers for professional development is sufficiently low-stakes. The risks associated with providing professional development to the wrong teacher manifest on the learning of the students who have teachers who should have been identified as needing professional development. Further, such a system would not be very cost-effective as funds for professional development would be wasted on teachers who do not need it. Finally, it may not be possible to identify the particular training teachers need based solely on TVAs as no information is provided about the areas in which teachers are deficit. By combining the TVA with other measures of teacher effectiveness, these risks and costs can be minimized as well.

An alternative that may be even more cost-effective and provide more timely information to administrators on teachers' needs is to use the TVA estimates to identify teachers for more robust data collection. For example, an evaluation system that calls for teachers to be observed using a rubric 10 times in one year may be prohibitively expensive to implement system-wide. However, a system that calls for teachers who are found to be ineffective using TVAs to be identified for additional data collection may enable administrators to identify those teachers who may be in need of additional supports, thus enabling the targeted allocation of professional development funds. This approach may also enable administrators to ascertain whether the evaluation system provided inaccurate data. This approach also solves a key problem associated with TVAs, which is that they do not provide any information regarding teachers' areas of deficit, which would instead be collected using the observation rubric or students' responses to questions about teachers' practices.

Whether they are intended for high stakes or low stakes decisions, the evaluation system should be designed to minimize unintended consequences. For example, the flaws inherent in TVA estimates of teacher effectiveness should be communicated to principals and superintendents responsible for using the information provided by the evaluation system such that they understand the limitations of the TVA method. Whether the state evaluation system requires that TVAs be used, enthusiastic administrators may use these estimates of their own volition to make hiring and firing decisions. The evaluation system may create the conditions for high stakes uses simply by providing the estimates to principals and superintendents. Training and documentation regarding the advantages and disadvantages of each approach used in the evaluation system should be provided. Further, the recommendations above

should help guard against the potential for teachers to be shifted around by grade level or subject to avoid having a TVA estimate.

Implications for Further Research

The implications for future research go beyond the direct implications of this study of classroom effects to the broader picture of how one conducts assessments of VAM performance. The main rationales for continuing to assess and compare VAMs, at a time when these methods are being deployed in state evaluation systems as teacher performance criteria, are to determine the extent to which there may be harmful unintended consequences for students, and the extent to which the evaluation systems promote higher student achievement, learning, and later life outcomes. Many of these considerations will have implications for the well-being of teachers as well. Taylor & Turner (2011), for example, demonstrated that an observation-based teacher evaluation system deployed in Cincinnati public schools, promoted student achievement both during the year following teacher evaluation and in long-term follow-up; however, no such study that I am aware of has been attempted for a TVA-based evaluation system or one consisting of both TVA and observations. The questions for scholars of education policy go beyond examining the virtues of each model to demonstrating how the harmful impacts of these imperfect models can be minimized and their advantages maximized, such that teachers can be evaluated fairly. As these systems are deployed, new opportunities will arise for studying the teacher labor market and the academic outcomes of students.

The central challenges of conducting research into VAMs are the trustworthiness of the models and the transparency of the methods used to make these assessments. A straightforward implication of this study is that the methods may not be sufficiently

trustworthy, and that additional methods for minimizing the disturbances from contemporaneous influences operating in the classroom should be tested. None of the methods demonstrated that they could adequately minimize the influence of this disturbance to be considered as essential for incorporation into TVA evaluation systems.

Transparency is critical as findings regarding the trustworthiness of the models, like the models themselves, must be understood by laypersons who may distrust highly technical methods that cannot be easily understood. I propose that this study has contributed to education scholars' understanding of how to assess the trustworthiness of VAMs in a transparent manner. Generally, the use of simulation data to examine the trustworthiness of VAMs may not be sufficiently transparent for lay audiences. Simulations are useful because they get around the problem of not having a true teacher effect to which the TVAs can be compared. The true teacher effects, which the TVAs are purported to measure, are unobserved (or equivalently, latent or missing). The simulations allow researchers to generate the "true" teacher estimate. Despite these advantages, the fact that the data are "made up" remains a challenge to their transparency and resulting efforts to examine the trustworthiness of the VAMs. The stylized nature of the data and the number of assumptions that must be employed in generating the data are subjective decisions that are then added to further subjective decisions about the choice of VAMs to examine and the methods by which these VAMs should be compared.

The credibility of these simulations can be improved by reducing the subjectivity of this decision process. This can be done, for example, by calibrating the inputs to real data (*e.g.*, Schochet & Chiang, 2011); by comparing the performance of the models on both simulated and real data (*e.g.*, McCaffrey et al., 2004); or, as in the present study, by limiting

the investigation to a single disturbance at a time. Under the last approach, further disturbances that add realism can be examined in robustness checks. The disturbances due to this violation are clarified by not being confounded by additional noise, and the limitations can be more easily communicated to skeptical audiences.

Second, the unconfoundedness assumption (though not SUTVA) can be examined in other more conventional settings. For example, researchers can develop randomized trials, much like the Tennessee STAR study, whereby teachers and students in multiple settings are randomly assigned to each other (Nye, Konstantopolous & Hedges, 2004). As noted previously, under these circumstances the average performance of each student would constitute a measure of teacher effectiveness unconfounded by student and teacher assignment. Consequently, this average could be assumed to be the teacher's true effectiveness, and the TVAs could be estimated and compared to this measure. This would not imply that the estimates satisfied the SUTVA assumption, as SUTVA violations can occur under any assignment pattern. Although non-experimental approaches cannot be assumed to satisfy the assumption of unconfoundedness, they can approximate this to the extent that credible natural experiments exist. These natural experiments can lend further support to investigations into the trustworthiness of VAMs for estimating TVAs (*e.g.*, Jacob, 2011), serving for example as replication studies of more expensive and onerous randomized trials. The advantage of conducting new randomized trials, rather than relying on existing trial data, is that effectiveness can be examined in the context of a deployed VAM-based teacher evaluation framework.

Third, a conventional approach in psychometric studies is to use cross-instrument validation techniques. Such techniques can be applied here. The more measurements taken

on a process, the more reliably the process will be understood (for example, see Willett, Singer & Martin, 1998). Under a comparison design that includes multiple methods, the TVA estimates would be compared to teacher effectiveness estimates obtained from observation rubrics (Pianta et al., 2008; Schochet, 2011), teacher surveys (Rowan, Correnti & Miller, 2002), student surveys (Kahle, Meece & Scantlebury, 2000; Koth, Bradshaw & Leaf, 2008) or combinations of these approaches. In states such as North Carolina, where teacher evaluation systems are being developed consisting of multiple approaches, opportunities exist for comparisons to be made. Although none of the instruments used in the comparison purport to represent a "true" teacher effect, multiple approaches that similarly rank teachers have a low probability of emerging from chance occurrence, and the influence of outliers in this evaluation process can be identified.

Finally, the means by which these models are compared should be as transparent as possible. Although no scholar doubts the importance of rank correlations and mean squared errors, there are some problems with using these methods to compare VAM performance and translate them to a lay audience. First, these methods may not be sensitive enough to detect minor statistical but high clinical differences between VAMs. In this study, I used the mean squared error and Spearman rank order correlations as "statistical" methods of comparison. Under both methods, these models outperform what we typically see when using these statistics in empirical frameworks. Secondly, however, these statistics lack practical meaning to lay audiences including policymakers who would benefit from understanding the differences between these models in terms of the number of teachers likely to be affected by the errors implied by these statistical measures. Therefore, in this study, I have demonstrated the comparison using a set of "practical" measures, including the level of agreement that

each model had with the true effect in ranking teachers in the bottom 5%, and the count of the number of teachers who would be affected by the choice of a lower-performing model over a higher performing one. These are gross person-level manifestations of VAMs' trustworthiness, representing the movement of teachers around a predetermined cutoff in a distribution, which may have more practical meaning to policymakers.

Concluding Comments

Evaluation systems that include TVAs have already been developed and implemented in the United States (*e.g.*, in Florida and Texas) and thanks to Race to the Top, are disseminating further into states like North Carolina. These systems have strong political support that persists despite reservations of scholars and stakeholders about the impact these evaluation systems will have on the teacher workforce and student learning and achievement. Teacher evaluation may have a significant and lasting impact on student achievement (Taylor & Tyler, 2011). The method by which teachers are evaluated may be an important factor, however.

Although this study showed that the error in measurement from TVAs is much lower than what is considered safe in a research context for estimating population parameters, these errors may not be low enough to use the data to take action on individual teachers, and thus caution is warranted. As I have noted, however, it is not true to suggest that a state-level evaluation system based on TVAs will have a uni-directional negative effect on teachers' job security. In fact, these methods can provided better information to principals about teachers' effects on students' performance on state standardized tests than simple averages would provide. In addition, the data can be supported with information provided to principals regarding the flaws associated with TVAs, such as those illustrated in this study. Finally,

they are based on more objective criteria than either survey or observational data, and they measure productivity rather than inputs.

When the advantages and disadvantages of these multiple approaches are considered, an evaluation that relies partly on all of these methods may be an important tool for principals. In North Carolina, for example, the state is adding a VAM teacher measure to an evaluation system consisting of five observational criteria, and requiring that teachers who fail to meet a predetermined effectiveness cutoff after a certain number of years be targeted for dismissal. Even in a scenario where the evaluation system does not contain other teacher effectiveness measures, but may contain a probationary period for teachers identified as ineffective, principals can intervene by using the TVA data formatively to identify teachers for further data collection by other methods, identifying those teachers who need professional development support, and providing that support.

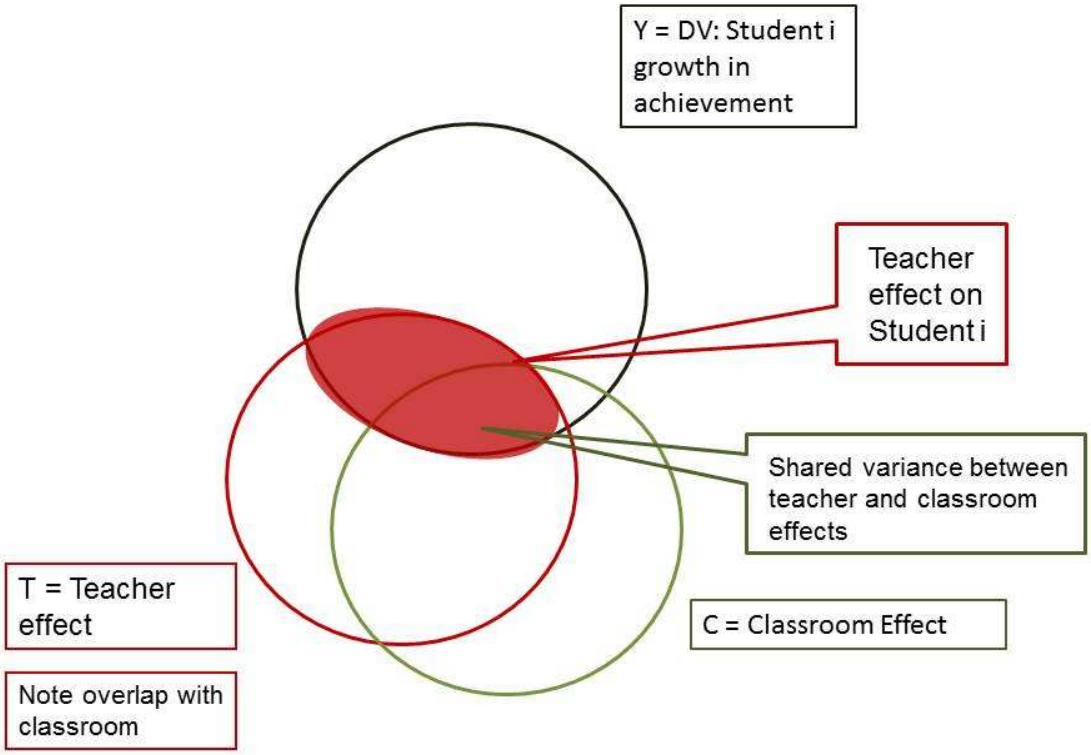

Figure 1.1. Shared variance between student growth and classroom and teacher effects.

Table 1.1.

Summary of Models

	Type	**Cross-Sectional or Panel**	**Time Invariant Covariates**	**Lagged Outcomes (Pretests)**	**Teacher Effect Parameter**	**School Random Effect**
HLM2	Nested random effects	Cross-sectional	Yes	1 in same subject	Teacher random effect (EB shrinkage estimator)	No
HLM3				1 in same subject		Yes
HLM3+				2 in each of 2 subjects		
SFE	Fixed effects	Panel	Differenced to zero	None (all outcomes as DV)	Mean of within-teacher residuals	N/A
TFE		Cross-sectional	Yes	1 in same subject	Teacher fixed effect (dummy variable)	
SFEIV		Panel (2 periods)	Differenced to zero	once lagged as endogenous predictor;	Mean of within-teacher residuals	
TFEIV				twice lagged as instrument (same subject)	Teacher fixed effect (dummy variable)	
DOLS				1 in same subject		
URM	Hybrid RE & FE	Cross-sectional	No	2 in each of 2 subjects used to calculate composite	Teacher random effect (EB shrinkage estimator)	No

Table 1.2.

List of Variables in Actual NC Data used in Calibrations and HLM4 Analysis

Student Variables	Description	Correlated Covariate Calibration	HLM4
asian	Asian Ethnicity	x	x
black	African American Race	x	x
multi	Multiracial	x	x
hispanic	Hispanic Ethnicity	x	x
amindian	American Indian	x	x
male	Gender = Boy	x	x
frlnch*	Free lunch	x	x
redlnch*	Reduced price lunch	x	x
islep*	Currently is LEP	x	x
waslep*	Was LEP in previous year	x	x
ex_dis	Disabled	x	x
ex_aig	Gifted	x	x
movediy*	Moved during school year		x
movedpy*	Moved between school years		x
movedpymiss	Missing, moved between school years		x
underage	Student is under age for grade	x	x
overage	Student is over age for grade	x	x
force_move	Change of school due to advancement	x	x
ma_peer/rd_peer	Avg of peers' standardized exam scores, previous year	x	x
Classroom Variables			
num_students	Number of students in classroom	x	x
ma_peer_sd / rd_peer_sd	Classroom SD of previous year's exam scores	x	x
lep75	Class LEP above 75th percentile	x	x
dis75	Class disabled above 75th percentile	x	x
aig75	Class gifted above 75th percentile	x	x
frl75	Class FRL above 75th percentile	x	x
oage75	Class overage above 75th percentile	x	x
Teacher Variables			
teach_exp	Teacher Experience (yrs)	x	
elem_ed_ic	Licensed elementary education: initial/continuing	x	

any_cont	Any type of continuing license	x	
any_init	Any initial license	x	
any_cont_tested	Any type of continuing license in a tested subject	x	
any_init_tested	Any type of initial license in a tested subject	x	
School Variables			
pctfrpl	School level percent free/reduced pric lunch	x	x
asian_mean	Proportion Asian	x	x
black_mean	Proportion Black	x	x
amindian_mean	Proportion American Indian	x	x
hispanic_mean	Proportion Hispanic	x	x
multi_mean	Proportion Multiracial	x	x
tot_ppx	Total per pupil expenditures	x	x
actper1k_l1	Violent acts per 1,000 previous year	x	x
stsrate_l1	Short term suspension rate previous year	x	x
adm	Average daily membership (Student pop.)	x	x
admsq	Squared ADM	x	

Table 1.3

Calibration of Variance Component Inputs

	Unconditional		**Conditional**	
Level	**Math**	**Reading**	**Math**	**Reading**
Teacher	0.134	0.152	0.017	0.012
School	0.138	0.120	0.034	0.016
Residual (student)	0.452	0.310	0.352	0.243
Teacher proportion	0.185	0.261	0.042	0.045
Student proportion	0.625	0.533	0.874	0.894
School proportion	0.190	0.206	0.085	0.060

Table 1.4

Mean Squared Error, Variance Decomposition Models

	Proportion of Variance at Classroom Level						
	0	0.02	0.04	0.06	0.08	0.1	0.12
URM	0.014	0.022	0.030	0.038	0.046	0.054	0.061
HLM3P	0.011	0.018	0.025	0.032	0.039	0.046	0.052
SFE	0.022	0.023	0.025	0.027	0.028	0.030	0.032
DOLS	0.026	0.034	0.041	0.049	0.057	0.065	0.073

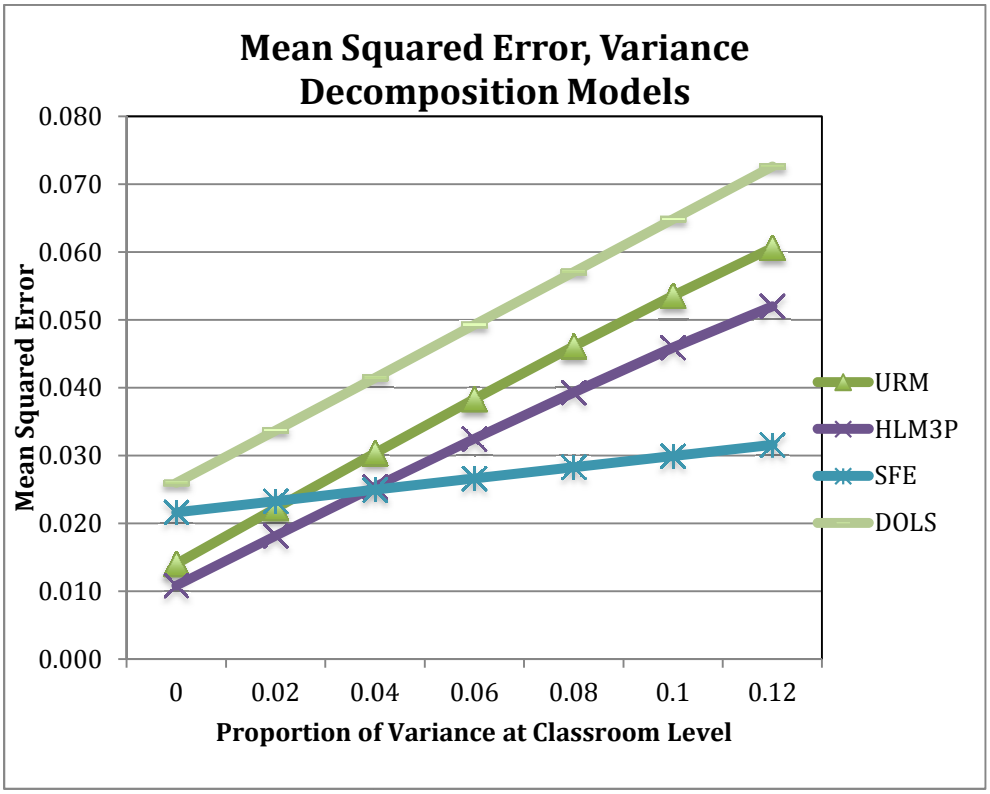

Figure 1.2. Mean squared error, variance decomposition model.

Table 1.5

Correlation of Estimated Teacher Effect with True Effect at Varying Levels of Classroom Variance

	Proportion of Variance at Classroom Level						
	0	0.02	0.04	0.06	0.08	0.1	0.12
URM	0.946	0.905	0.856	0.795	0.718	0.613	0.456
HLM3+	0.955	0.913	0.864	0.802	0.724	0.618	0.460
SFE	0.941	0.900	0.851	0.791	0.714	0.610	0.454
DOLS	0.909	0.869	0.822	0.763	0.689	0.589	0.437

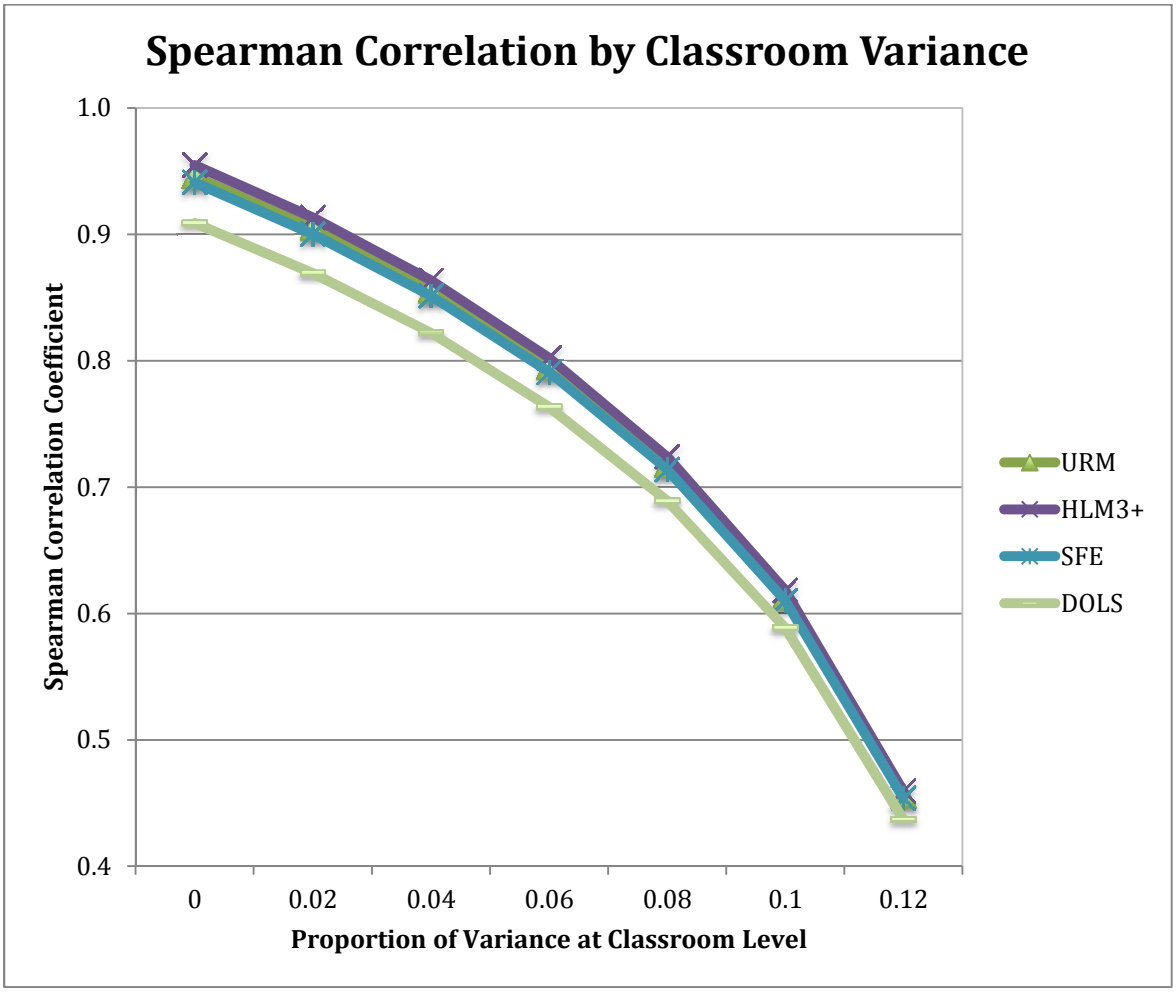

Figure 1.3. Spearman correlation by classroom variance.

Table 1.6

Percent Agreement between TVA and True Teacher Effect on Identifying Bottom 5% of Teachers

	Proportion of Variance at Classroom Level						
	0	0.02	0.04	0.06	0.08	0.1	0.12
URM	97.443	96.564	95.813	95.035	94.226	93.354	92.372
HLM3	97.712	96.812	96.014	95.167	94.331	93.472	92.439
SFE	97.333	96.487	95.750	94.994	94.170	93.313	92.351
DOLS	96.622	96.036	95.352	94.725	93.990	93.172	92.218

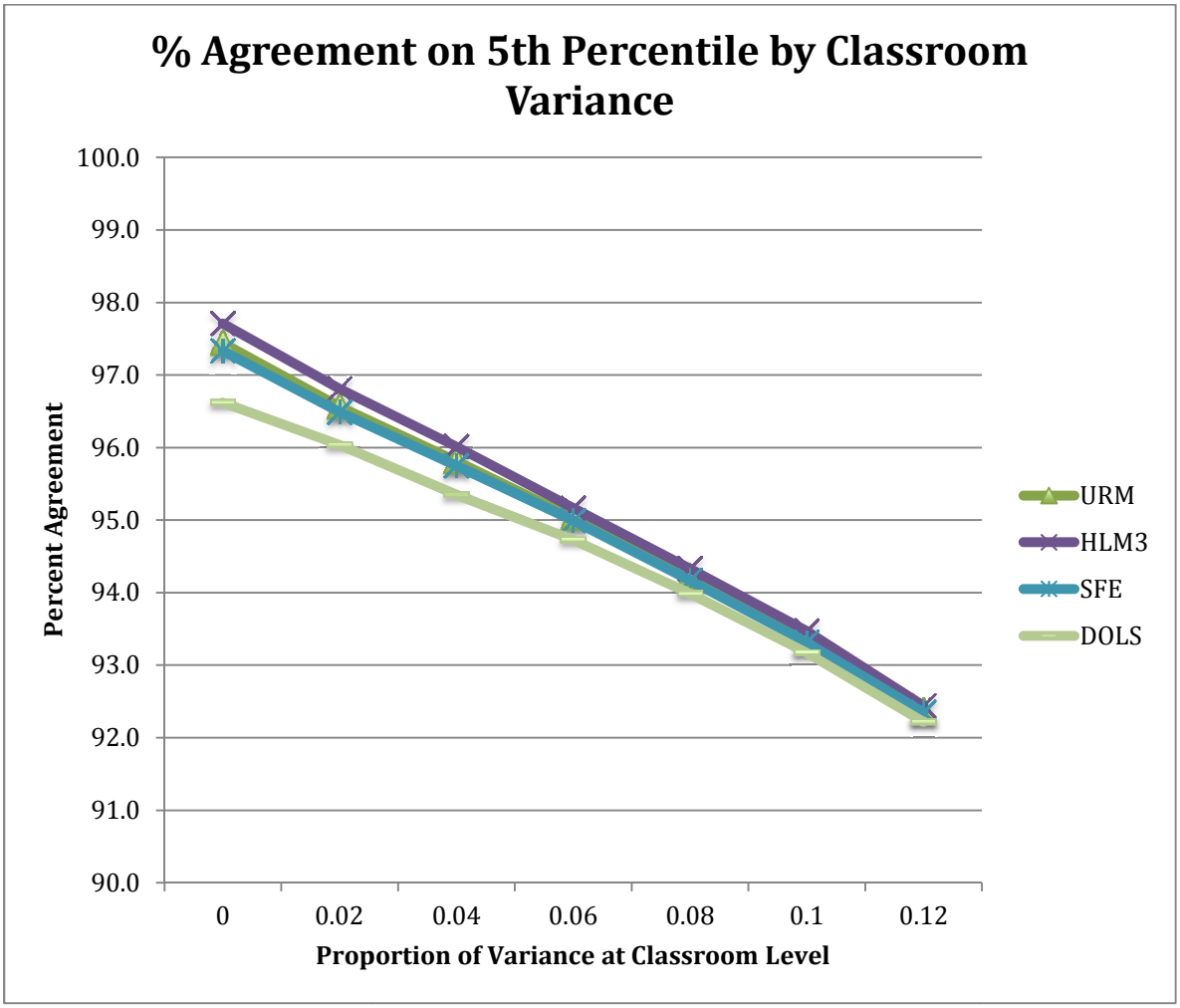

Figure 1.4. Percent agreement on 5th percentile by classroom variance.

Table 1.7

Number and Proportion of Teachers Falsely Identified as Ineffective by Classroom Proportion of Variance

	0		**0.02**		**0.04**		**0.06**		**0.08**		**0.10**		**0.12**	
	Prop.	**N**	**Prop.**	**N**	**Prop.**	**N**	**Prop.**	**N**	**Prop.**	**N**	**Prop.**	**N**	**Prop.**	**N**
URM	0.013	118	0.017	155	0.021	193	0.025	225	0.029	264	0.034	309	0.040	357
HLM3+	0.012	108	0.016	144	0.020	182	0.024	219	0.028	256	0.034	302	0.039	352
SFE	0.014	122	0.018	160	0.022	194	0.025	227	0.029	263	0.034	309	0.040	358
DOLS	0.017	153	0.020	180	0.023	210	0.027	241	0.030	274	0.035	312	0.039	355

Table 1.8

True Scores of Teachers Falsely Identified as Ineffective

	Proportion of Variance at Classroom Level						
	0	0.02	0.04	0.06	0.08	0.1	0.12
URM	-1.41	-1.33	-1.24	-1.14	-1.02	-0.85	-0.61
HLM3+	-1.43	-1.35	-1.26	-1.16	-1.03	-0.86	-0.63
SFE	-1.40	-1.32	-1.23	-1.13	-1.01	-0.84	-0.61
DOLS	-1.33	-1.26	-1.18	-1.09	-0.98	-0.81	-0.58

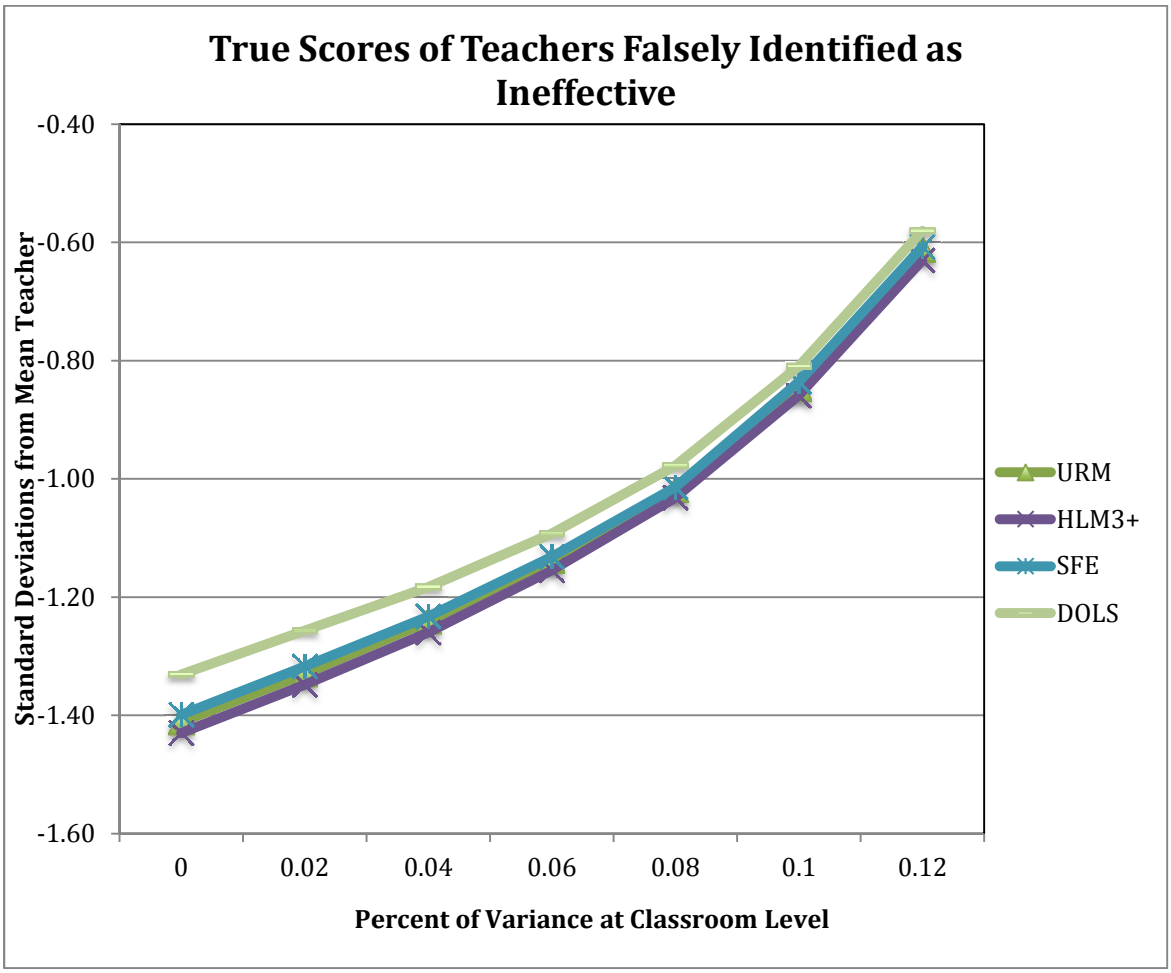

Figure 1.5. True scores of teachers falsely identified as ineffective.

Table 1.9

Mean Squared Error

	Correlation between Teacher and Classroom Effects						
	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60
URM	0.08	0.08	0.08	0.08	0.08	0.07	0.07
HLM3P	0.06	0.06	0.05	0.05	0.06	0.05	0.05
SFE	0.11	0.10	0.10	0.09	0.09	0.09	0.08
DOLS	0.14	0.14	0.13	0.13	0.13	0.14	0.13

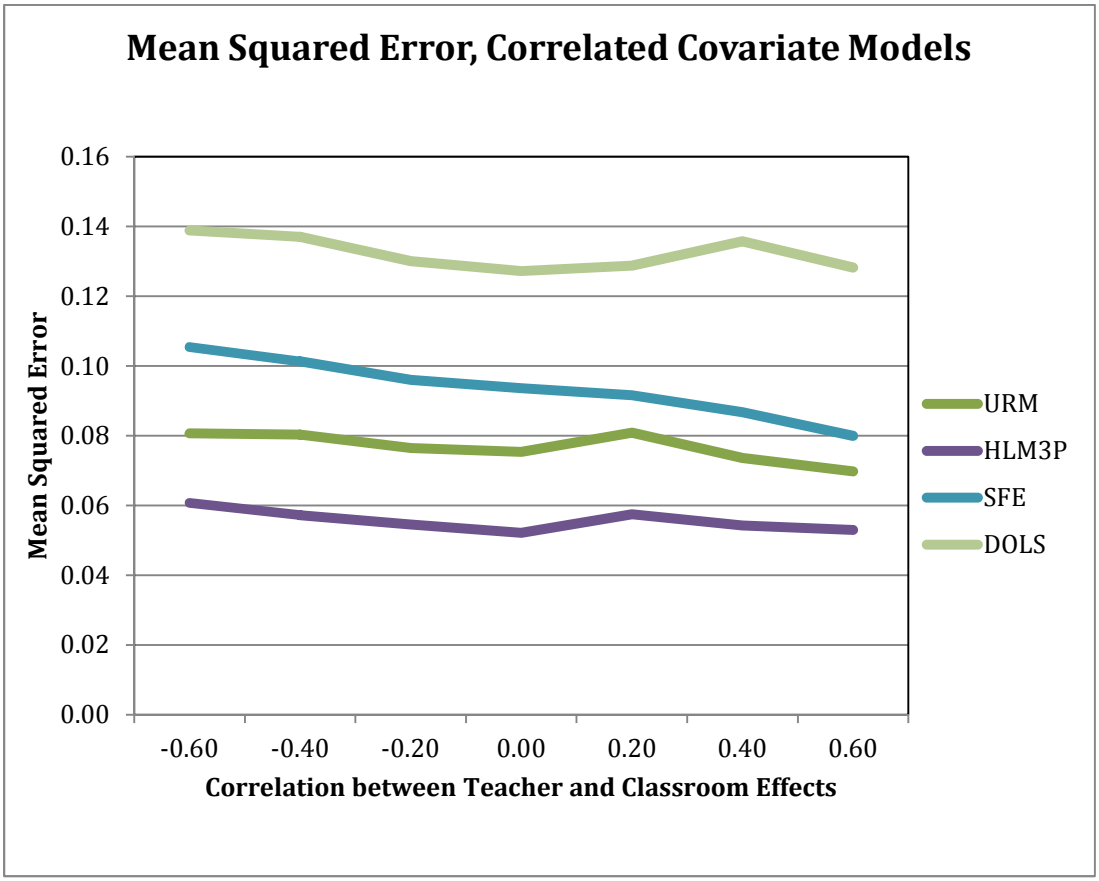

Figure 1.6. Mean squared error, correlated covariate models.

Table 1.10

Spearman Rank Order Correlation by Correlation between Teacher and Classroom

	Correlation between Teacher and Classroom						
	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60
URM	0.670	0.668	0.686	0.701	0.684	0.716	0.728
HLM3+	0.758	0.767	0.783	0.788	0.760	0.777	0.780
SFE	0.612	0.619	0.636	0.639	0.634	0.657	0.674
DOLS	0.684	0.681	0.699	0.701	0.698	0.688	0.704

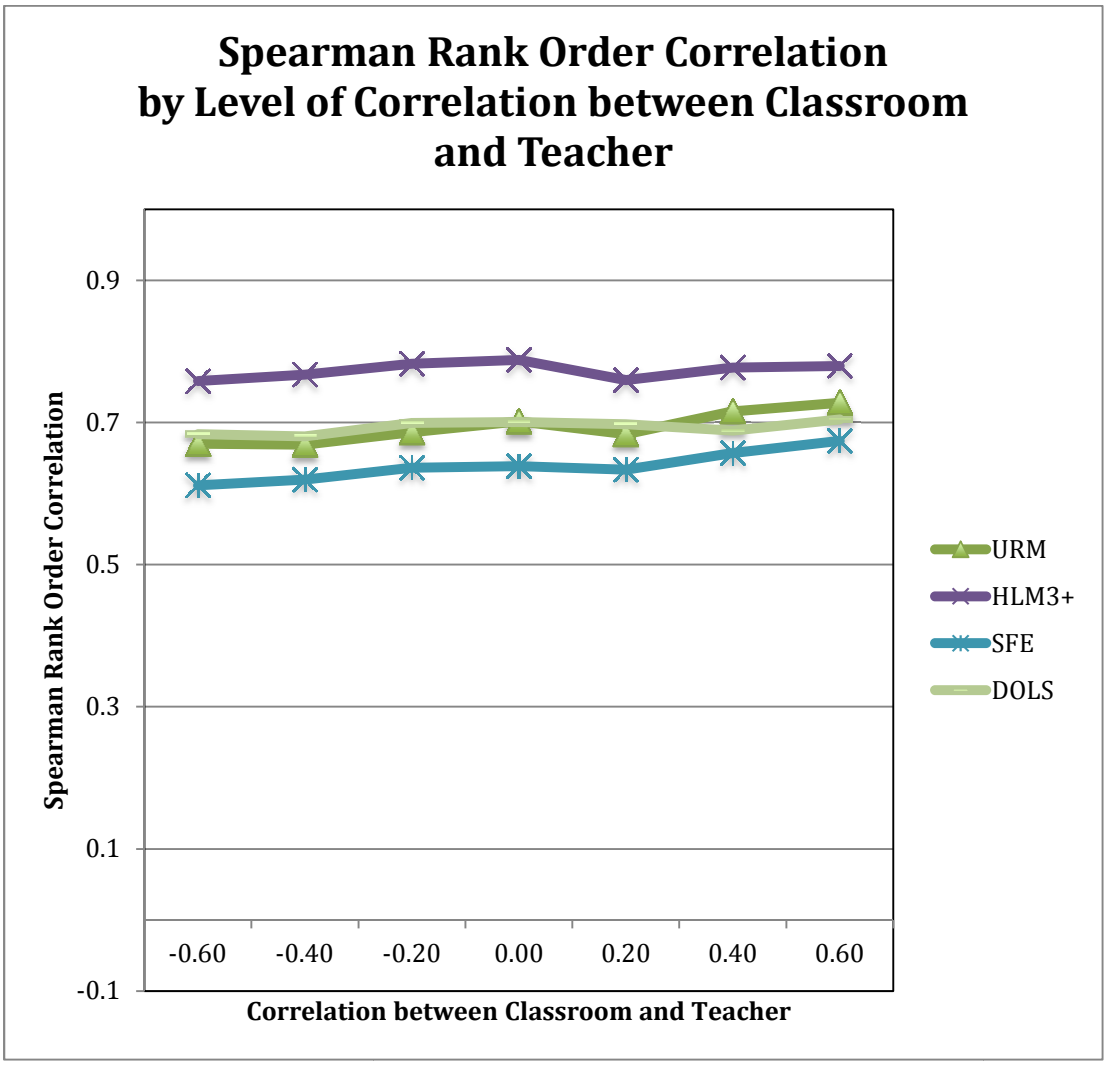

Figure 1.7. Spearman rank order correlation by level of correlation between classroom and teacher.

Table 1.11

Percent Agreement between TVA and True Estimate on Bottom 5% of Teachers

	Correlation between Teachers and Classrooms						
	-0.6	-0.4	-0.2	0	0.2	0.4	0.6
URM	93.8	93.7	93.9	93.9	94.0	94.3	94.4
HLM3+	94.7	94.7	94.8	94.8	94.7	94.9	94.9
SFE	93.4	93.3	93.5	93.4	93.5	93.8	93.8
DOLS	94.0	93.9	94.0	93.9	93.9	94.0	94.0

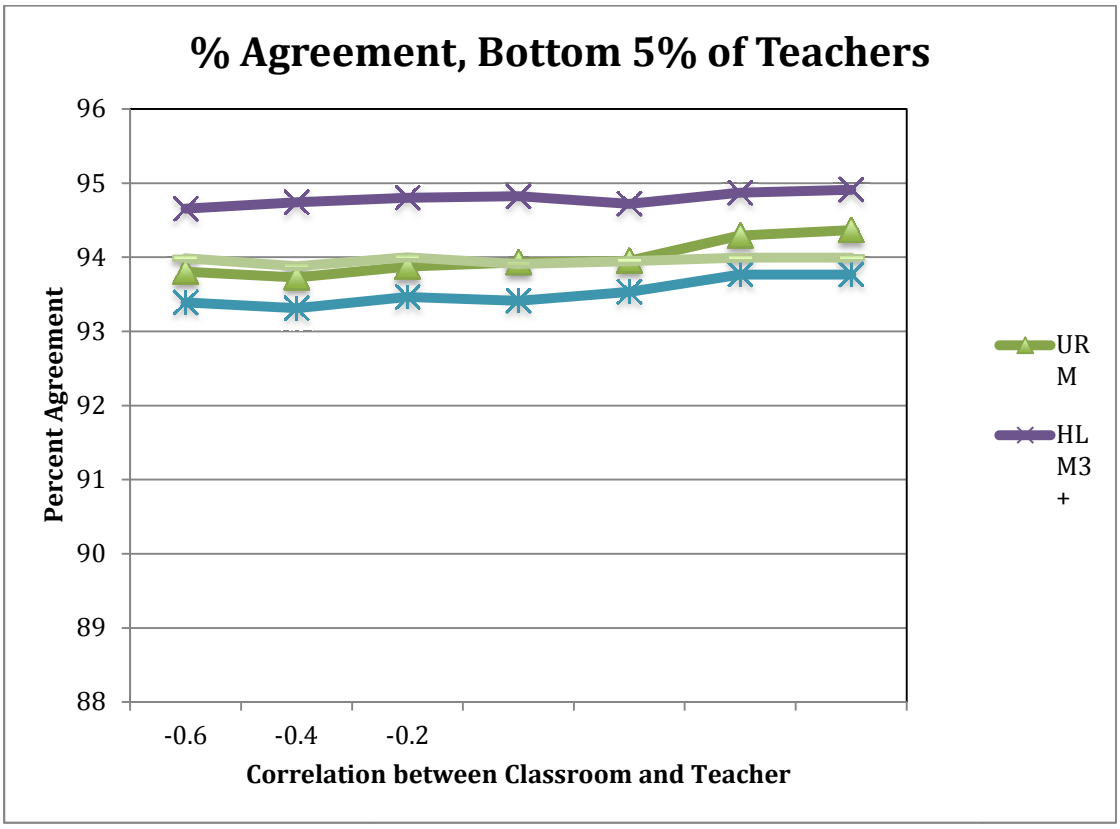

Figure 1.8. Percent agreement, bottom 5% of teachers.

Table 1.12

Proportion and Number of Teachers Falsely Identified as Ineffective

	Correlation between Teacher and Classroom Effects													
	-0.60		**-0.40**		**-0.20**		**0.00**		**0.20**		**0.40**		**0.60**	
	Prop.	**N**	**Prop.**	**N**	**Prop.**	**N**	**Prop.**	**N**	**Prop.**	**N**	**Prop.**	**N**	**Prop.**	**N**
URM	0.031	277	0.032	288	0.030	270	0.031	277	0.030	273	0.030	267	0.028	256
HLM3+	0.027	241	0.027	245	0.026	232	0.026	232	0.026	236	0.026	231	0.025	229
SFE	0.033	300	0.034	306	0.033	293	0.033	293	0.033	293	0.032	288	0.030	273
DOLS	0.030	270	0.031	282	0.029	265	0.031	275	0.030	274	0.031	275	0.030	269

Assuming 9,000 Fifth Grade Teachers

Table 1.13

True Z Scores of Teachers Falsely Identified as Ineffective

	Correlation between Teacher and Classroom Effects						
	-0.60	**-0.40**	**-0.20**	**0.00**	**0.20**	**0.40**	**0.60**
URM	-0.93	-0.93	-0.95	-0.96	-0.98	-1.00	-1.03
HLM3+	-1.08	-1.09	-1.10	-1.12	-1.12	-1.12	-1.12
SFE	-0.85	-0.86	-0.86	-0.89	-0.92	-0.92	-0.96
DOLS	-0.95	-0.96	-0.99	-0.99	-0.98	-0.98	-1.00

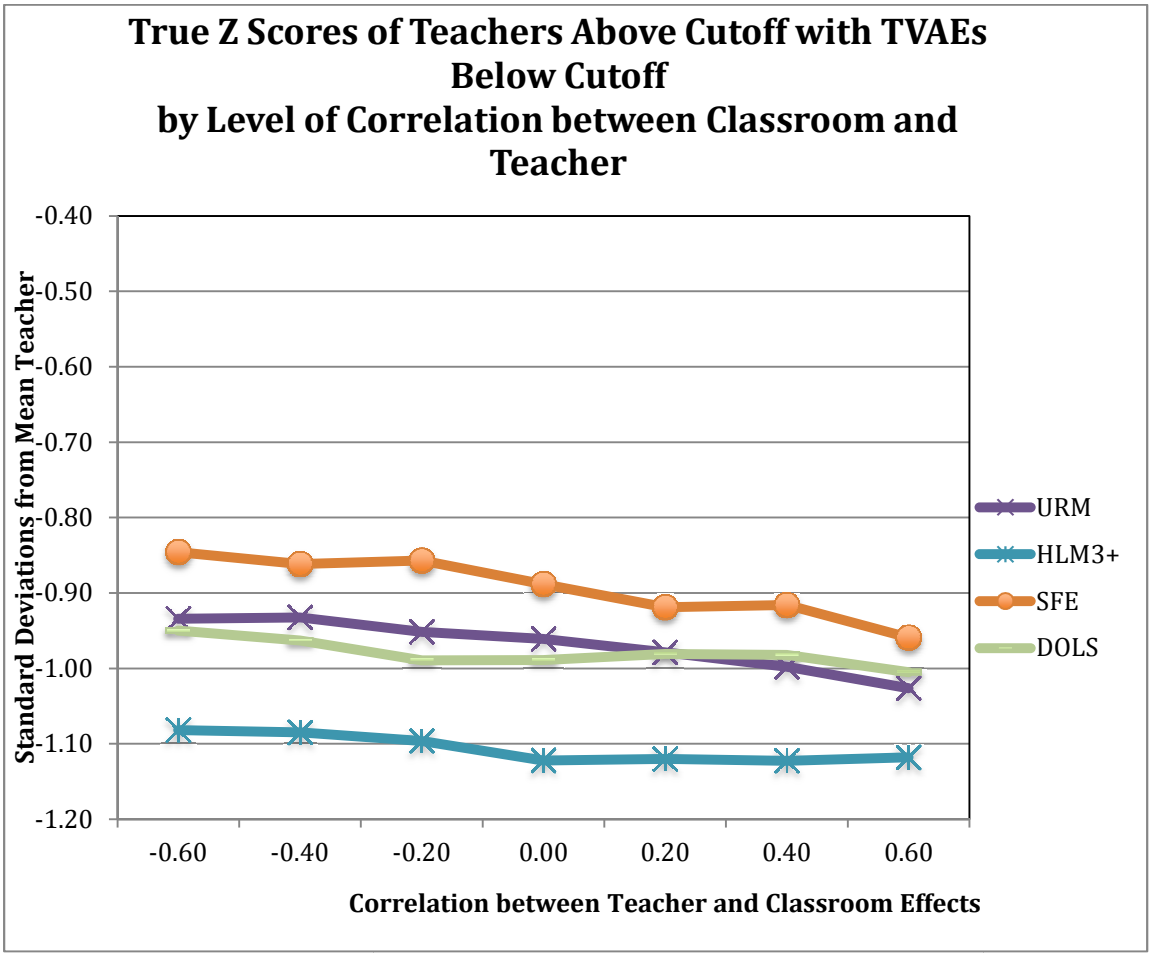

Figure 1.9. True Z scores of teachers above cutoff with TVAEs below cutoff by level of correlation between classroom and teacher.

Table 1.14

Correlation of Estimated Teacher Effect with True Effect at Varying Levels of Classroom Variance

4 Cohorts							
	Proportion of Variance at Classroom Level						
	0	0.02	0.04	0.06	0.08	0.1	0.12
URM	0.947	0.906	0.857	0.796	0.719	0.614	0.456
HLM3+	0.955	0.914	0.864	0.803	0.724	0.619	0.460
SFE	0.942	0.901	0.852	0.792	0.715	0.611	0.454
DOLS	0.911	0.872	0.825	0.766	0.691	0.590	0.439

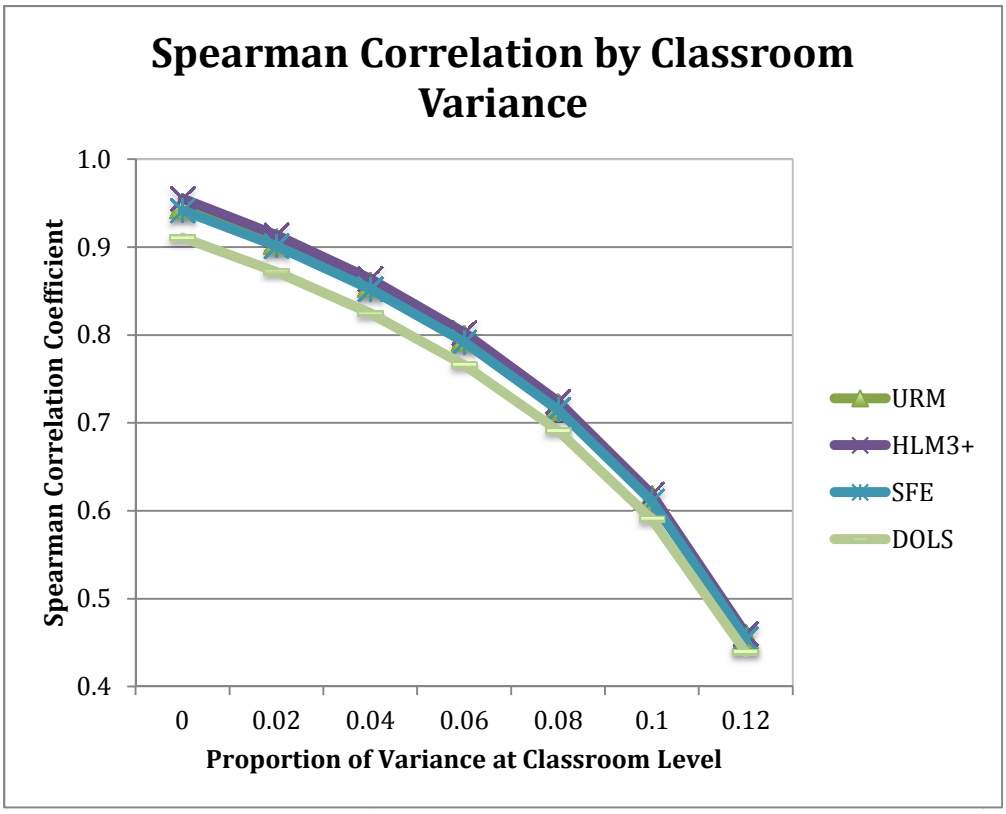

Figure 1.10. Spearman correlation by classroom variance.

Table 1.15

Percent Agreement between TVA and True Teacher Effect on Identifying Bottom 5% of Teachers

4 Cohorts							
	Proportion of Variance at Classroom Level						
	0	0.02	0.04	0.06	0.08	0.1	0.12
URM	97.400	96.612	95.839	95.049	94.199	93.333	92.379
HLM3	97.570	96.739	95.923	95.148	94.276	93.390	92.420
SFE	97.285	96.492	95.758	95.001	94.170	93.316	92.372
DOLS	96.677	96.110	95.429	94.749	93.969	93.193	92.267

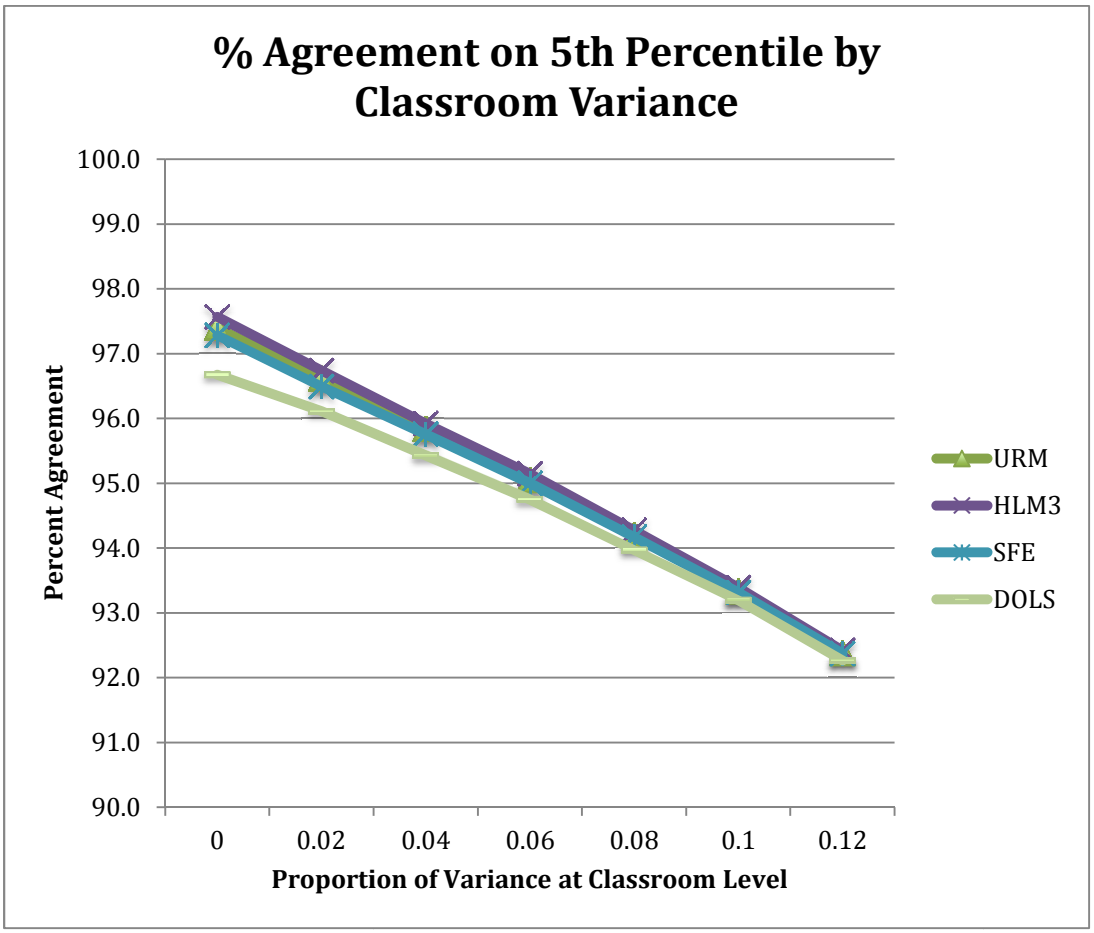

Figure 1.11. Percent agreement on 5th percentile by classroom variance.

Table 1.16

Number and Proportion of Teachers Falsely Identified as Ineffective by Classroom Proportion of Variance

4 Cohorts														
	0		**0.02**		**0.04**		**0.06**		**0.08**		**0.10**		**0.12**	
	Prop.	**N**	**Prop.**	**N**	**Prop.**	**N**	**Prop.**	**N**	**Prop.**	**N**	**Prop.**	**N**	**Prop.**	**N**
URM	0.007	62	0.008	74	0.010	88	0.011	103	0.013	120	0.015	136	0.018	160
HLM3+	0.006	57	0.008	69	0.009	85	0.011	99	0.013	115	0.015	137	0.017	157
SFE	0.007	65	0.009	77	0.010	89	0.011	103	0.013	118	0.015	137	0.018	161
DOLS	0.008	72	0.010	86	0.011	98	0.012	110	0.014	124	0.016	140	0.018	161

Table 1.17

True Scores of Teachers Falsely Identified as Ineffective

4 Cohorts							
	Proportion of Variance at Classroom Level						
	0	0.02	0.04	0.06	0.08	0.1	0.12
URM	-1.74	-1.68	-1.61	-1.53	-1.42	-1.27	-0.99
HLM3+	-1.77	-1.71	-1.62	-1.54	-1.43	-1.25	-1.00
SFE	-1.73	-1.66	-1.59	-1.52	-1.41	-1.27	-0.99
DOLS	-1.67	-1.60	-1.55	-1.47	-1.38	-1.22	-0.96

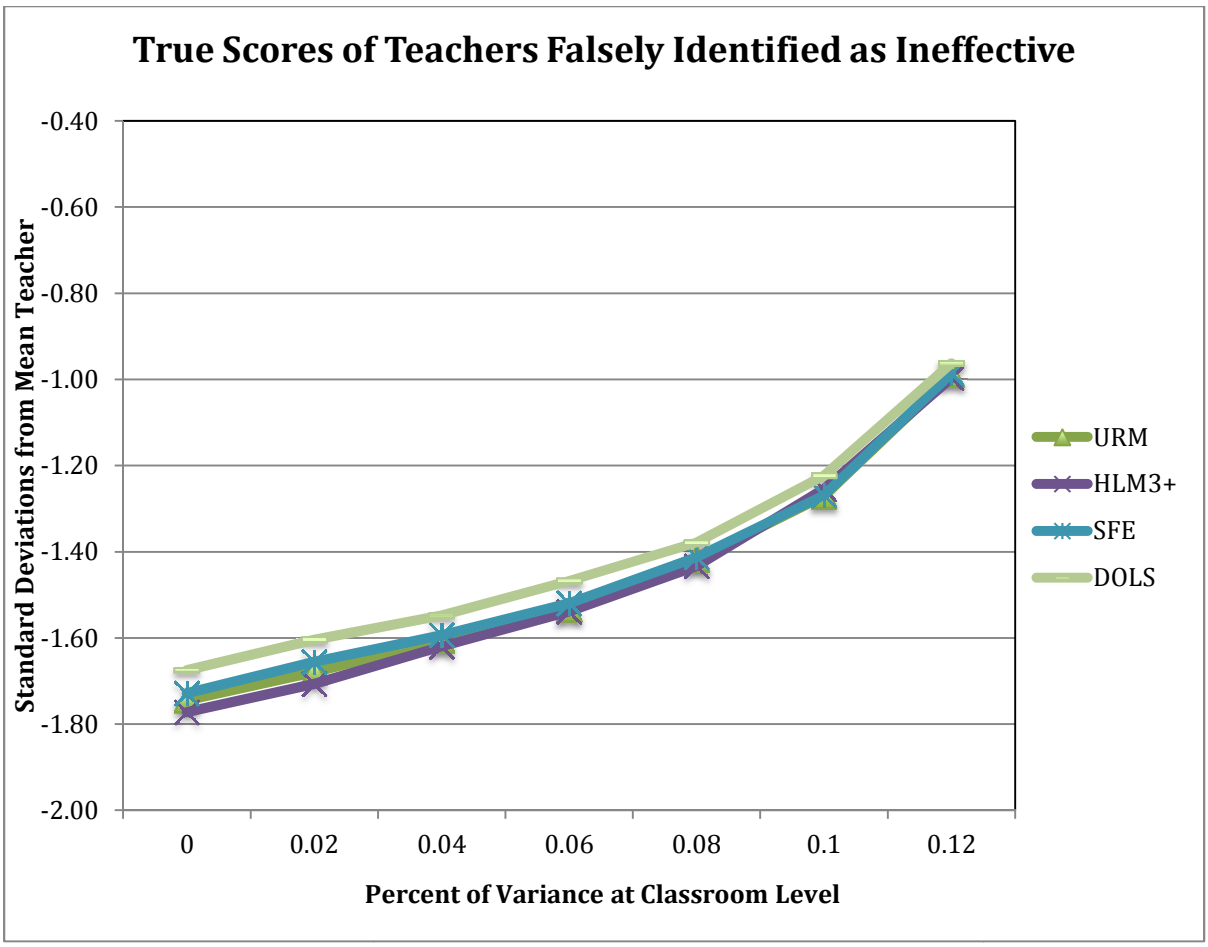

Figure 1.12. True scores of teachers falsely identified as ineffective.

Table 1.18

Correlated Covariate Simulation: 2 vs. 4 Cohorts

					True Scores, Proportion and Number of Teachers Falsely Identified as Ineffective					
	Spearman Rank Order Correlation		**Percent Agreement on Bottom 5%**		**2 Cohorts**			**4 Cohorts**		
	2 Cohorts	**4 Cohorts**	**2 Cohorts**	**4 Cohorts**	**True Score**	**Proportion**	**No. of Teachers**	**True Score**	**Proportion**	**No. of Teachers**
URM	0.686	0.748	93.87	94.50	-0.961	0.032	284	-1.059	0.028	249
HLM3+	0.762	0.820	94.68	95.32	-1.079	0.027	245	-1.181	0.024	212
SFE	0.638	0.706	93.49	94.07	-0.896	0.033	299	-0.987	0.029	264
DOLS	0.697	0.748	94.05	94.45	-0.966	0.030	270	-1.063	0.028	252

At 0.20 correlation between teachers and classrooms

Table 1.19

HLM4 Model Results: Variance Decomposition Models

	Proportion of Variance at Classroom Level				
Spearman Rank Order Correlations	0.864	0.803	0.725	0.620	0.461
Percent Agreement, Bottom 5%	96.02	95.17	94.35	93.47	92.41
Teachers Falsely Identified as Ineffective					
Mean True Z-Score	-1.26	-1.15	-1.04	-0.87	-0.63
Proportion Falsely Identified	0.02	0.02	0.03	0.03	0.04
N (Assuming 9,000 5th Grade Teachers)	180.86	218.25	254.77	301.44	352.55

Table 1.20

True Scores, Proportion, and Number of Teachers Falsely Identified as Ineffective (HLM4 and CCOV Models)

HLM4	Spearman Rank Order	Percent Agreement on Bottom 5%	True Score	Proportion	N (of 9,000)
-0.6	0.726	94.31	-1.01	0.028	256
-0.4	0.720	94.30	-1.00	0.029	265
-0.2	0.730	94.30	-1.01	0.028	254
0	0.728	94.142	-1.02	0.029	263
0.2	0.707	94.10	-1.02	0.029	259
0.4	0.730	94.36	-1.03	0.029	259
0.6	0.740	94.46	-1.05	0.028	249

CCOV	Spearman Rank Order	Percent Agreement on Bottom 5%	True Score	Proportion	N (of 9,000)
-0.6	0.802	95.23	-1.14	0.024	220
-0.4	0.802	95.08	-1.14	0.025	221
-0.2	0.800	95.20	-1.14	0.024	217
0	0.799	94.93	-1.14	0.025	224
0.2	0.763	94.62	-1.09	0.026	237
0.4	0.775	94.69	-1.10	0.027	240
0.6	0.763	94.66	-1.10	0.027	243

CHAPTER 2. CAREER RELEVANT INSTRUCTION AS A CAUSAL PROCESS IN THE CAREERSTART TEACHING INTERVENTION PROGRAM

Teacher effects on student achievement are widely seen in education policy as large, suggesting a potential for student gains by focusing policies on intervening with teachers (Nye, Konstantopolous, & Hedges, 2004). Education policy researchers have identified several policy interventions for promoting teacher effectiveness: high stakes accountability consisting of rewards or sanctions, promoted largely by economists (*e.g.*, Gordon, Kain & Staiger, 2006); pre-service training, promoted mainly by education scholars (*e.g.*, Hill et al., 2008; Wayne & Youngs, 2003); and *in-service* training or professional development (PD) in the form of interventions offered to teachers to alter their instructional practices, promoted by scholars across a spectrum of disciplines including policy, education and social work (e.g., Grossman et al., 2000; Orthner et al., 2010). Ultimately, all three approaches imply that teachers will adopt more effective teaching strategies though the routes to adoption differ (Hiebert & Morris, 2012).

A strong case has been made for high stakes accountability policies (e.g., Gordon, Kain & Staiger, 2006) in part because the evidence suggests that the relationship between training (e.g., certification programs) and effective practice may be weak (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2008; Goldhaber & Brewer, 2000; Harris & Sass, 2011; Smith, Desimone & Ueno, 2005). However, by examining easily-measured structural features of teachers' preparation (e.g., the number of content area courses required by a preparation program), teachers' pre-preparation characteristics (e.g., college entrance exam scores), the

effects of specific teacher practices may not be realized (Kennedy, 2010; Palardy & Rumberger, 2008). Alternatively, these practices should be observed and measured (Rowan, Correnti & Miller, 2002; Smith, Desimone & Ueno, 2005), and effective practices incorporated into pre-service and in-service training and PD (Palardy & Rumberger, 2008). *Evidence-based teaching practices* oriented around promoting teacher effectiveness and ultimately student achievement can be identified using rigorous designs such as experiments or rigorous non-experimental methods (Cohen, Raudenbush & Ball, 2003; Rowan, Correnti & Miller, 2002).

Rigorous studies informing best practices should build on the evidence demonstrating the processes by which children learn (Brophy & Good, 1986; Cohen, Raudenbush & Ball, 2003; Hill et al., 2008; Hill, Rowan & Ball, 2005). A well-known objective process of student learning is school engagement, which represents an emotional, behavioral, or cognitive investment in learning tasks (Finn, 1989; Fredricks, Blumenfeld & Paris, 2004). Motivation has been shown to be a subjective precursor of engagement (Eccles & Wigfield, 2002), and studies suggest that teaching strategies that target motivation may alter engagement (Orthner, Jones-Sanpei, Akos & Rose, 2012; Wentzel & Wigfield, 2007). One practice strategy that has demonstrated some promise and has been shown to operate through known motivational processes is relevance, the practice of associating course material with affectively and cognitively important aspects of students' current or future lives (Hulleman, Godes, Hendricks & Harackiewicz, 2010). That is, course material is presented in ways in which it is associated with potential uses outside of the classroom. This may address several motivational challenges facing students in a typical classroom setting, where instruction is

typically decontextualized (e.g., Cordova & Lepper, 1996; Husman, Derryberry, Crowson & Lomax, 2004).

One relevance-based teaching practice intervention, CareerStart, supports teachers in their efforts to use career-relevant instruction (CRI)—the relevance of jobs and careers that students might have in the future—as a middle school motivational technique (Orthner et al., 2010). A randomized control trial (RCT) of CareerStart demonstrated significant effects on teacher use of CRI in math and science (Rose, Woolley, Orthner, Akos, Jones-Sanpei, 2012); and in student outcomes such as school valuing, a measure of emotional engagement with schooling (Orthner et al., 2012), middle school end-of-grade math achievement (Woolley et al., in press), and high school end-of-course biology achievement and progress towards graduation (Woolley, Unick, Rose & Orthner, in revision).

However, the process by which CareerStart achieves these effects has only been partially confirmed. Understanding this process is essential to its further development and dissemination as a best practice. A logic model for CareerStart (Figure 2.1) suggests that by increasing teachers' use of CRI (Rose et al., 2012). CareerStart promotes higher engagement and achievement partly through the effect of CRI on students' motivation. Although the RCT enabled causal inferences about CareerStart from correlational analyses, the same cannot be said for causal inferences about CRI. That is, CRI was not randomly assigned to teachers. Instead, teachers who participated in CareerStart had a higher probability of using CRI—itself a causal inference due to the randomization of CareerStart—but other factors not under the control of the evaluator, such as teachers' years of experience, which are also associated with achievement, were shown to have influenced teachers' use of CRI (Rose et al., 2012).

This study uses a rigorous econometric method typically used in non-experimental settings, instrumental variable estimation (IVE), combined with the random assignment of CareerStart to estimate a local average treatment effect (LATE; Angrist, Imbens & Rubin, 1996) for CRI and demonstrate a causal effect of CRI on engagement and achievement despite the non-randomness inherent in teachers' use of CRI. By establishing CRI as the pathway through which CareerStart promotes achievement, this study has potential to both support the logic model for the intervention and CareerStart as a validated best practice for teaching.

In the following sections, I discuss the roles of engagement and motivation in student learning. I then describe five theories of motivation that may inform the importance of relevance as an instructional practice strategy, review the extant literature demonstrating the congruence of these theories, suggesting multiple pathways for relevance. I then review studies suggesting the instructional strategies that relevance may ultimately change. I follow this with a discussion of career relevant instruction and the CareerStart treatment program. Following that, I describe the methods and results of an investigation, based on an experimental evaluation of CareerStart, showing that career relevant instruction is a process for the benefits promoted by CareerStart and has a causal impact on students' engagement and achievement. Practice and policy implications for the findings are provided.

Engagement and Motivation

The processes by which children learn may inform practice strategies that can be investigated for evidence in promoting achievement (Brophy & Good, 1986; Cohen, Raudenbush & Ball, 2003; Hill et al., 2008; Hill, Rowan & Ball, 2005). School engagement is a well-known objective process of student learning, which represents a cognitive

investment in learning tasks (Finn, 1989; Fredricks, Blumenfeld & Paris, 2004) and can be targeted for teaching practice intervention (Wentzel & Wigfield, 2007).

School Engagement as a Learning Process

School engagement, an objective dimension implying an individual commitment or investment of cognitive resources in learning tasks, has long been considered a causal process in learning and an important precursor to student achievement (e.g., Finn, 1989; Finn & Rock, 1997). Engagement is now understood to consist of three sub-dimensions—behavioral, emotional and cognitive (Fredricks, Blumenfeld & Paris, 2004): a) behavioral engagement concerns student conduct and demonstration of staying on-task; b) emotional engagement concerns students' attitudes, interests and values towards schooling; and c) cognitive engagement, which is a relatively new type of engagement, relates to motivational goals and self-regulating behavior.

Engagement is associated with students' successful transition from middle to high school (Anderman, Maehr & Midgley, 1999), and is a predictor of school drop out, with less engaged children more likely to drop out of school (Finn, 1989). Because academic resilience is partly explained by engagement, disengagement is a risk factor for school failure, and a potential target for intervention (Finn & Rock, 1997). Disengagement from school, which is often driven by factors outside of the control of the children or the school, such as parental work and income or neighborhood safety, is a developmental process that begins at an early age (Finn, 1989). A lack of fit between developmental needs and the social opportunities and structures provided to middle school adolescents may be an underlying source of maladaptive behaviors of children undergoing intense biological and social changes (Roeser, Eccles & Sameroff, 2000). For example, a key part of students' developmental growth is to recognize

the link between their academic behavior and their performance in schools (Roeser, Eccles & Sameroff, 2000), and low performers are less likely than others to have recognized such connections (Anderman et al., 2010). To give these students the same opportunities as their better prepared peers, developmentally appropriate engagement interventions are needed, and introducing such interventions in the middle grades may act as a protective investment for later engagement in high school and prevent drop out (Balfanz, Herzog & Mac Iver, 2007; Janosz, Archambault, Morizt & Pagani, 2008). Again, because disengagement is often predicated on factors that are external to school and not under the direct control of educators, an important strategy is to strengthen the resilience of students' engagement to these distractions and obstacles (Balfanz, Herzog & Mac Iver, 2007).

Motivational Teaching Practices

A goal of improving engagement suggests a teaching practice method that strengthens its causal precursors. Motivation, a subjective rather than objective antecedent of student achievement, may be a leverage point for intervening in student engagement (Wentzel & Wigfield, 2007). Motivation has been defined as "the process by which goal-directed activity is instigated and sustained" (Schunk, Pintrich, & Meece, 2008, p. 4), or a "set of processes that provides energy for different behaviors" (Wentzel & Wigfield, 2007, p. 262). In education research, the study of motivation is largely conducted by educational psychologists with the goal of understanding "how to motivate people to engage in new learning" (Weiner, 1990, p. 618). Motivation is distinct from engagement but a direct causal antecedent, and those students reporting greater motivation demonstrate greater engagement (Balfanz, Herzog & Mac Iver, 2007). Thus, if motivation is improved via targeted teaching practices that are applied alongside quality content expertise, engagement may improve as well.

Opportunities exist for the implementation of motivational teaching practices. Two historical tendencies in instructional practice present significant challenges for motivating students to learn. First, teachers present new material in highly decontextualized forms to ensure it has the highest generalizability, and in doing so, it robs the material of what may make it interesting to students; paradoxically, any context that may be interesting may be a useful way to motivate (Cordova & Lepper, 1996). Second, school is by its very nature future-oriented, in that children are schooled to prepare them for their futures. This suggests three additional problems. First, early adolescents are by this developmental stage primarily focused on the present. Second, if the future is not seen as positive and hopeful, it may be an insufficient motivation, which may be one reason why minority students relatively underperform (Phalet, Andriessen & Lens, 2004). Third, if the future goals cannot be translated into more proximal near-term goals, then they may seem too far out of focus or reach to be realistic motivators, a challenge that may be even more urgent for younger children (Husman & Lens, 1999). By disconnecting the skills and knowledge of the classroom from the factors that give them relevance, less-prepared learners who do not inherently see these links are put at disadvantage. Thus, a candidate for motivational intervention is to imbue the content with relevance to children's lives. In the next section, I discuss the motivational theories that may lend support to this hypothesis.

Relevance and Motivation

Relevance describes students' perception of one or more potential uses for the material that they learn in school, outside of just successfully completing their schoolwork. Individual or multiple but congruent motivational theories may explain the process by which relevance benefits student motivation, engagement, and therefore learning (Hulleman, Durik,

Schweigert & Harackiewicz, 2008). These include expectancy value theory (Eccles & Wigfield, 2002); interest theories (Hidi & Harackiewicz, 2000; Schiefele, 1991); goal orientation theories (Ames, 1992; Harackiewicz, Barron, Pintrich, Elliot & Thrash, 2002); self-determination theory and intrinsic and extrinsic motivation theories (Lepper & Cordova, 1992; Ryan & Deci, 2000); and identity or possible future selves theory (Markus & Nurius, 1986).

Motivational Theories

First, the task value component of expectancy value theory (EVT) makes explicit that the utility of what is being learned may provide important cognitive and affective components to the motivation to learn (Eccles & Wigfield, 2002). Task value in EVT consists of four perceptive constructs associating a task with some relative value: attainment value (the importance of doing well on a task), intrinsic value (enjoyment from undertaking a task), utility value (the use of the task to complete other goals), and cost belief (the opportunity cost of what is lost by undertaking an activity) (Wigfield & Eccles, 1992). The third of these, utility value, describes the perception of future uses for these tasks, which may be a mediator for relevance (Wigfield & Eccles, 1992).

Second, relevance may be directed through an effect on students' goal orientations. Goals are generally dichotomized into mastery and performance goals (Ames, 1992), though a further distinction can be made between performance-approach and performance-avoid (Harackiewicz, Barron, Pintrich, Elliot & Thrash, 2002). A mastery goal orientation is marked by a student's perception that the content is useful in and of itself and so worth learning, but not just to demonstrate performance relative to others (Ames, 1992). Performance-approach is an orientation concerning performance relative to others that is

marked by a student's perception of the importance of doing well on a task relative to others on a task; alternatively, performance-avoid is an orientation away from tasks that the student perceives that they will not perform well at relative to other students (Harackiewicz, Barron, Pintrich, Elliot & Thrash, 2002). Relevance may not operate through a performance orientation. Evidence shows, however, that the learning environment becomes more oriented towards performance goals than mastery goals during middle school (Anderman & Midgely, 1997). Reorienting the middle school classroom environment to support the relevance of learning tasks may therefore support a mastery goal orientation (Cordova & Lepper, 1996).

Third, interest theories (Schiefele, 1991) support the notion that students' affect towards certain subjects can be a powerful motivator and can be manipulated within the classroom setting. As noted previously, interest is a component of task value in EVT (Wigfield & Eccles, 1992), but it has received substantial attention in its own right as a motivational construct. The key facet of interest is that it is an emotional or affective constituent of motivation, not a cognitive one (Schiefele, 1991). Relevance may help teachers associate course material with interests that students already have, or it may help teachers cultivate new interests in students.

Fourth, intrinsic and extrinsic motivation theories (Lepper & Cordova, 1992), and theories of self-determination (Ryan & Deci, 2000), suggest that students have basic needs for autonomy and self-competence, such that motivation for a specific task that is initiated as an internal state (intrinsic) will be sustained more deeply and for longer than motivation for the same task that is initiated by another person (extrinsic). That is, cognitive motivation can be promoted and maintained if students feel like they are pushing themselves rather than being coerced by others (Ryan & Deci, 2000). For students to be autonomous and make

informed choices regarding the direction of their work, they should have information about how the tasks and material can benefit them in some way outside of the classroom (Skinner & Belmont, 1993). Extrinsic and intrinsic motivation for a specific task were once thought of as extreme ends of a single dimension, but are now viewed as two negatively correlated dimensions in which extrinsic rewards may undermine intrinsic motivations (Lepper, Corpus & Iyengar, 2005). Intrinsic interest in the material may make it self-sustaining, reducing the need to use motivation-undermining extrinsic methods (Hidi & Renninger, 2006).

Fifth, identity theories of motivation, notably possible selves theory, suggest that students' positive images of their future selves can be called on to compel students to behave in accordance with these images and act as a motivator for learning (Markus & Nurius, 1986; Oyserman, 2008). Generally, identity motives, which students may not be consciously aware of, exert pressure to seek desirable identities and avoid undesirable identities (Vignoles, Golledge, Regalia, Manzi & Scabini, 2006). Consequently, instruction that is relevant to desirable future possibilities can provide cues to students to learn new concepts that are consistent with these identities, and students who report that their possible selves appear clearer to them perform better academically (Oyserman, 2008; Oyserman, Bybee & Terry, 2006).

Multiple Motivational Pathways for Relevance

As suggested in the preceding discussion, studies support the relationship between relevance and each of five key motivational constructs. Alternatively, relevance may operate concurrently or sequentially through groups of these constructs. The extant literature examining multiple motivational constructs lends support to the suggestion that many of these theories are congruent. In this framework, students' intrinsic interest in the material is,

in the context of relevance as a practice strategy, the ultimate motivational construct preceding engagement, with the other constructs causally preceding and supporting intrinsic interest.

First, as noted previously, associating the tasks or content with aspects of value to students may activate an affective response and promote interest in the material (Cordova & Lepper, 1996). Interest and intrinsic motivation are often used interchangeably (Hidi & Harackiewicz, 2000), and in fact, they may share some precursors, *e.g.*, need-related feelings may underly the development of interest and intrinsic motivation (Krapp, 2005). But interest should be seen as a distinct validated causal antecedent of intrinsic motivation (Hidi & Harackiewicz, 2000), and in fact is likely to be mediated by other factors preceding intrinsic interest, including task value and autonomy (Hulleman, Durik, Schweigert & Harackiewicz, 2008).

Second, a mastery goal orientation may be promoted by associating the content or tasks with factors of value to students' current or future lives (Husman & Lens, 1999; Meece, 2003). Goal orientation and interest may also have a relationship; for example, mastery goals are associated with increased interest among low-achieving children (Harackiewicz & Elliot, 1993). The role of relevance in goal development is supported by a related theory, future time perspective (FTP; Husman & Lens, 1999). FTP suggests that near-term self-regulation can be influenced by future goals if these future goals promote proximal subgoals that can help attain the future goals (Miller & Brickman, 2004). Motivation is thus improved by informing the potential future use of a task or of knowledge, and by providing sufficient information about the authenticity of the task or knowledge (Husman, Derryberry, Growson & Lomax, 2004).

Third, task value may follow goal orientation and interests, but may also be independently promoted via relevance as relevance directly clarifies the current or future utility of a task. The task value component of EVT has both a direct effect on intrinsic or extrinsic motivation, as well as an indirect effect mediated by autonomy (Hidi & Harackiewicz, 2000). Identity or possible selves theory also suggests the importance of task value as a related concept in the context of relevance (Orthner et al., 2010). Relevance, by promoting the utility of a task can provide a framework for students to link the task to their identity formation and clarify those identities. Finally, situational interest in the task, and goals developed via the knowledge associated with utility of the task, may also spur the development of deeper interests, consistent with the four-phase model proposed for interest development (Hidi & Renninger, 2006).

These constituents of the relevance-motivation process form the basis for the knowledge and values that motivates students and encourages them to engage in learning. In the next section, I discuss relevance interventions, which target motivational processes with the ultimate aim of promoting engagement and performance.

Relevance as an Intervention in Instructional Practice

A body of research dating back to Keller (1987) suggests that improving students' perceptions of the relevance of learning tasks and schoolwork is a method to directly intervene with students in a cognitive and affective manner. Students begin their schooling reporting a high level of motivation and demonstrating high engagement, but their motivation is subsequently dulled by the schooling experience, and evidence across all levels of schooling shows that to change this, authentic academic work consisting of meaningful problems that are relevant to students may be needed (Marks, 2000). Decontextualized

instruction, which I now turn to discuss, may be the primary teaching practice that interrupts students' motivation.

The Challenges of Decontextualized Instruction

Generally, content is taught by teachers in the most decontextualized form possible, which serves to maximize its generalizability to a wide array of possible uses (Cordova & Lepper, 1996). However, researchers have identified several problems with this approach. First, it removes from the material the information about the utility of the task or knowledge that might encourage students to learn the material, dulling students' autonomy. Second, the inherent future-bias of these uses, which is consistent with the purpose of schooling, means that the utility of the material being taught is not immediately relevant to students even if such utility can be made explicit. As a result, goal formulation may be challenged. Third, it removes from the material the context that makes it interesting, which may serve as a primary affective motivator that acts independently to motivate students' interest in the material.

Information about task utility. To generalize the material being learned to a wide number of uses, educators have typically presented new tasks and content without the context for their uses (Cordova & Lepper, 1996). This is an admirable objective but is highly idealized, implicating the assumption that students themselves will visualize at present or in the future how the tasks and content may be of value to them. More realistically, decontextualization of the material, by separating the material from the information with which the material is associated, does not accord with how young children are shown to learn (Cordova & Lepper, 1996). By leaving the impetus for connecting the content and tasks to their current or future uses to the students themselves puts many students at a disadvantage

and leaves these students without valuable, if imperfect or incomplete, information about the utility of the tasks and content (Assor, Kaplan & Roth, 2002; Greene et al., 2004).

Addressing this may be as simple as having the teacher explain how the learning task will help the students achieve certain goals, and asking the students about how they feel about such tasks in this regard (Assor, Kaplan & Roth, 2002). Providing the context for the material will directly influence the perceived instrumentality of a task, implying that it is an immediate outcome of relevant classroom instruction (Greene et al., 2004). In this context, relevance has been included as either a component of autonomy support (Assor, Kaplan & Roth, 2002; Skinner & Belmont, 1993) or as a component of motivating tasks (Greene et al., 2004), both of which are viewed as causal precursors of autonomy. Assor, Kaplan & Roth (2002), argued that their findings suggest that autonomy, or choice, is ineffective without the support from the teacher, in the form of understanding the relevance of the material in a way needed to act autonomously. These authors entered relevance into their model as a component of autonomy support or enhancement. In Greene et al. (2004), relevance was included in the motivating tasks dimension, not autonomy support; although autonomy support did not predict the perceived instrumentality or utility of a task, motivating tasks did (Greene et al., 2004). In both cases, relevance was part of a broader measure predicting autonomy.

Future Relevance for the task. Dewey argued against tying content to future use, arguing that it sent a message devaluing students' current interests (Dewey, 1916). But schooling is inherently future oriented, largely lacking links to the present interests of students (Hagay & Baram-Tsabari, 2011). Therefore a focus on children's futures may often be lacking, which may be one reason why minority students who often perceive lesser future

prospects underperform, may also reduce the motivation of some groups of children particularly those already at risk for school failure, which may contribute to the achievement gap (Phalet, Andriessen & Lens, 2004).

The efforts at making topics and tasks relevant as described in the previous section, if they can be associated with future uses, are intended to address this shortcoming. By making explicit the future adult uses for the learned tasks and material, it accomplishes two related objectives. First, it helps students to see the value of the task to their futures and thus illustrates the value of being successful at the task in the present; and it illuminates positive prospects for students' futures by relating current success at the task to future success at the task as an adult (Husman, Derryberry, Crowson & Lomax, 2004; Husman & Lens, 1999; Miller & Brickman, 2004). This finding generalizes to minority populations (Phalet, Andriessen & Lens, 2004).

Interest in the task. The highly decontextualized form of teaching, meant to convey material in as generalizable a manner as possible, robs the material of what may make it most interesting to students (Cordova & Lepper, 1996). That is, the content is present without the context that may engage students' affective response to it, making it dull and boring and resulting in a steady decline in intrinsic motivation from elementary through high school (Cordova & Lepper, 1996). Associating the material with a meaningful context, including about its use in students lives' at present and in the future, may provide a way to invigorate students' interest in the material. Tasks that are both cognitively demanding and relevant in the sense of being useful for students' needs promote higher levels of motivation (Hoyt, 2005) and engagement (Marks, 2000). Lessons that employ conceptual teaching strategies, such as those that demonstrate the utility of math to real life, may encourage higher cognitive

engagement than those that focus on rote procedures and memorization (Smith, Desimone & Ueno, 2005).

Evidence for Relevance Oriented Teaching Strategies

There is ample evidence in the education literature for the benefits that may come from intervening with relevance-oriented teaching practices to reverse the decontextualization of course content. I divide these findings into relevance for students' *current* and *future* lives.

Current. In a study of a fantasy-based computer-aided math learning tool, children completed math problems using one of five conditions—a non-fantasy control condition, and four progressively embellished fantasy conditions, including a generic fantasy condition, a personalized fantasy condition whereby the student was able to choose the particular fantasy, and two nearly-equivalent conditions that also allowed the students to select from instructionally-irrelevant features (Cordova & Lepper, 1996). In general, students learning from the fantasy oriented programs benefited relative to those in the control condition, but students in the most embellished condition improved more, used deeper task involvement, more complex operations, and learned more (Cordova & Lepper, 1996).

In a study of a science curriculum based around linking science to students' lives, which included inquiries such as why the students should wear a helmet while bicycling, demonstrated large effect sizes that were significant three years after the intervention, continuing into high school; students demonstrated higher order science thinking skills and narrowed a gender gap with boys closing the gap on girls (Geier et al., 2008). Two further studies of a relevance intervention in which undergraduate students were asked to provide a written explanation on how a technique that they had just learned would have relevance to

their lives was shown to increase perceptions of utility value and maintained situational interest in the material, particularly among those lowest in performance between pre and posttest; further, perceived utility value mediated the intervention's effect on situational interest, and there was an interaction between the intervention and performance expectations or prior performance on situational interest and utility value (Hulleman, Godes, Hendricks & Harackiewicz, 2010).

Non-experimental designs add further support and suggest a directed focus on a particular type of relevance. For example, a correlational study involving college students suggests that identity relevance—that the task at hand was central to an individual's self-perception of their identity—promoted motivation (Britt, 2005). This suggests that meaningful forms of relevance that relate schooling activities to their primary purpose—to prepare children for their future—may be as useful as helping students to understand the uses of the material in their current lives (e.g., Geier et al., 2008).

Future. Future relevance, if made more explicit may work via means that are closer to the intent of schooling, and help students see the relation between school now and what they value for the future (Husman & Lens, 1999). Perceived instrumentality is seen as a future version of utility value that may promote extrinsic motivation. However, if a future instrumental use for a task can be imbued with a present day task value by realizing that competence in the present can provide future benefits, it may be possible for it to become intrinsic (Husman & Lens, 1999). A study of undergraduates in a human resources training program suggests that greater study time was associated with higher instrumentality for the learned tasks, with the focus on the future rather than present utility value; further, this observed instrumentality was associated with intrinsic motivation (Husman, Derryberry,

Crowson & Lomax, 2004). Another study with younger children demonstrated that when future careers were seen as education dependent (e.g., computer scientists or plumbers, rather than basketball players), students invested more time and effort in schoolwork by almost eight-fold; the relevance of what they were doing to future earnings encouraged motivation (Destin & Oyserman, 2010).

Value itself may have a time dimension to its perception. Students with longer future time perspective may be more persistent in working towards goals and get more satisfaction from current goal-oriented activities; that is, the drop-off of students' valuing for an activity as its utility recedes into the future may be less for these students (Husman & Lens, 1999). A correlational study of 44 African American high school students using a correlational design demonstrated a significant association between engagement and self-reported future relevance of the coursework (Crumpton & Gregory, 2011). It may also be possible to intervene to mitigate the drop off in valuing for tasks showing future utility. Oyserman, Terry and Bybee (2002) show that a program that helps youth focus on or imagine themselves as successful adults, and then to tie these images to current school activity motivation, was associated with higher school bonding, concerns about school performance, strategies to attain more realistic goals, better attendance and, among African American students, fewer problem behaviors.

One type of future relevance that has demonstrated some promise as a means of promoting intrinsic motivation through interest, goals and autonomy processes is career relevance. Studies show that, despite not being old enough for vocational training, middle school students do consider college and career plans and identities (Akos, Konold & Niles, 2004), and must make curricular choices that will determine future educational opportunities

(Akos & Galassi, 2004). Thus, getting students to think about their futures in an informed manner may require explicit efforts to inform them of the links between what they do in school and potential future outcomes. By tying tasks and material learned in the present to the instrumentality or utility of the material for successfully performing the tasks related to specific jobs and careers, students may associate otherwise decontextualized material with positive feelings; may develop both distal and proximal goals to strive towards; and may make better choices regarding engagement in and completion of schoolwork (Husman, Derryberry, Crowson & Lomax, 2004; Husman & Lens, 1999). I now turn to discuss a promising instructional intervention that is based on career relevance, CareerStart.

CareerStart: A Career Relevant Instruction Intervention

CareerStart is a middle-school teaching practice intervention that relies on career relevant instruction (CRI) as the means by which tasks and lesson materials are imbued with meaning that conveys information about the importance of the material, its potential affective value, and a distal goal in the form of being competent in a task that may have some utility later in life. CareerStart includes a newsletter distributed to teachers that describes the connections between emerging jobs and careers and the lesson content that they teach. Further, it includes 10 lessons in each of the four core subject areas—math, language arts, science and social studies—for a total of 40 lessons per grade level over all three middle grade levels. CareerStart lessons were developed by North Carolina teachers, and reviewed by curriculum specialists, to be integrated into the North Carolina Standard Course of Study so as to not further burden teachers. As an in-service practice intervention, it represents a type of professional development whereby teachers learn by doing, with only a brief pre-use training session to discuss the conceptual underpinnings of the program and provide

examples. Further, it marries conceptual tools—the use of career-relevance to contextualize instruction—with practical strategies for the use of these tools in the form of the 10 lessons per subject area, a key component of successful PD interventions (Birman, Desimone, Porter & Garet, 2005; Grossman et al., 2000).

An example CareerStart lesson is included in Appendix 2; it is an eighth grade math lesson about using percent change calculation in a retail job. As the example illustrates, the jobs and careers that are used to contextualize the lesson material are not limited to those requiring a college degree. Many of the lessons do associate content with jobs and careers requiring higher education or vocational education. The jobs and careers that were selected were chosen with local labor market conditions in mind such that they were presented as realistic opportunities to children being raised in this labor market.

Evidence for CareerStart and CRI

CareerStart was developed, implemented and evaluated in a partnership with one district in North Carolina. Fourteen middle schools in this district were randomly assigned to the treatment conditions, with 7 schools receiving CareerStart and 7 schools being in the control condition. In the treatment condition, teachers were trained in the conceptual foundation of CRI, provided easy access to the lessons, which were provided online, and received consultation from a district-level representative. Further, principals in these schools were strongly encouraged by the superintendent and district to have their teachers use CRI in their classrooms. A logic model for CareerStart (Figure 2.1) supports CRI as a key process through which CareerStart operates to promote student engagement and academic achievement.

Teachers in the control condition were not barred from using CRI, and more than 50% of students in the control schools reported that they agreed or strongly agreed that their teachers frequently used career examples, which may have come from textbooks or from original content and not from CareerStart lessons (Rose et al., 2012). Accordingly, CRI, or the use of career examples in instruction, was considered a proximal outcome of the treatment in the logic model for CareerStart because it could be observed in both treatment and control schools and could be conceived of as a general and potentially effective teaching practice that CareerStart was intended to promote.

The literature described in the preceding sections supports the theory of change that CareerStart operates through CRI. However, the model also indicates that an alternative pathway for the CareerStart effect may be through teacher effectiveness. The argument supporting this alternative pathway is that teachers participating in CareerStart receive training on the "new 3 Rs of education"—rigor, relevance and relationships—including not only the importance of career relevance as a motivational technique (through, for example, expectancy value theory or possible selves theory; Orthner et al., 2010), but the importance of setting high expectations for students (i.e., academic press; Lee & Smith, 1999). Therefore, the training may in fact help teachers to become more effective in general.

Using the randomized design, key portions of the CareerStart logic model have been validated. For teachers, CareerStart has been demonstrated as promoting the use of CRI among teachers in math and science, including a stronger effect for teachers with more years of service (Rose et al., 2012). Further, a mixed method fidelity study showed that in three of the seven treatment schools (labeled "low fidelity"), teachers used well less than half of the lessons on average and that in contrast to teachers in a selected high fidelity school, who

reported modifying the lessons to fit their students' needs, teachers in a low fidelity school elected not to use the lesson when they perceived the fit to be poor (Phillippo & Rose, 2012). A qualitative study suggested that some teachers misunderstood the rationale for the program (confusing it with a vocational training program) and expressed some concern about the fit of the lessons to their students needs (Woolley, Rose, Mercado & Orthner, 2013).

For students, CareerStart demonstrated a causal effect on students' valuing of school (Orthner et al., 2012). In this study, positive and significant *associations* between students' perception of CRI and both valuing and engagement were also demonstrated. In a qualitative study some teachers reported that the lessons increased students' involvement in the classroom and promoted engaging classroom discussions (Woolley, Rose, Mercado & Orthner, 2013). Critically, Woolley et al. (in press) demonstrate that CareerStart affects the ultimate outcome of student achievement. CareerStart was demonstrated to have significantly increased students' end-of-grade math achievement in middle school, at 0.61 points on the EOG test scale; this is equivalent to one-quarter of a standard deviation effect size and about 1/3 the size of the annual rate of growth in math for the average student. Finally, CareerStart has lasting effects on children in high school, with students from CareerStart middle schools performing better on end-of-course biology tests and accruing credits towards graduation at a higher rate than students from control middle schools (Woolley, Unick, Rose & Orthner, under review). These findings, particularly those of Orthner et al. (2012) and Woolley et al. (in press) showing the positive effect of CareerStart on middle school outcomes, provide strong justification for understanding the role of the mediating processes described in the logic model.

Research questions. The evidence cited above suggests that CareerStart promoted use of CRI in math and science, student engagement and valuing, and math achievement. It is important to understand the process by which these effects emerged. Correlational analysis also showed an association between CRI and engagement and math achievement. An unanswered question in these analyses is whether CRI is the linking process tying the teacher effect (higher CRI) to student outcomes (engagement, valuing and achievement). That is, is CRI the causal process mediating the treatment effect of CareerStart and promoting engagement and achievement? The non-random assignment of CRI complicates the effort to ascribe causality to this relationship. There may be other factors, both measured and unmeasured, that compelled teachers to use more or less CRI. Such factors may have been inherent to the teacher, such as their gender, content area, or ethnicity (Rose et al., 2012), or they may have been in response to the needs or traits of the particular group of students being taught or unforeseen situations that emerged on a day-to-day basis (Woolley, Rose, Mercado & Orthner, 2013).

This is an unanswered question in the logic model for CareerStart, and an affirmative answer would not only lend support to this intervention but add further support to the body of research concerning relevance as a motivational instructional technique. In this study, I use rigorous measurement and analysis methods to address this gap, answering the following important questions about the process by which CareerStart may promote academic achievement: (1) Does CRI promote higher engagement and valuing in school; (2) Does CRI promote higher end-of-grade eighth grade math and reading achievement; and (3) do the data support an alternative pathway through academic press, as measured by *rigor* (e.g., setting high standards for success) and *preparation for further schooling* (providing the support that

students need to achieve that success) as suggested by the logic model? Because of limitations in the data, I use eighth grade outcomes only. This approach deviates slightly from the longitudinal analysis used in Orthner et al. (2012) and Woolley et al. (in press) in which all three middle level grades were examined as outcomes of the treatment.

Method

Design

CareerStart was developed, implemented and evaluated in a randomized control trial (RCT) in 14 middle schools in a district in North Carolina. In the treatment condition (N = 7), teachers were given a brief training in the conceptual foundation of CRI; provided access to pre-packaged CareerStart lessons via a web site; and received consultation from the district. A causal estimate of teacher CRI and an effort to link CareerStart, CRI and the outcomes has not been attempted due to two challenges: non-random assignment of CRI and measurement of CRI by students' perceptions. I now turn to discuss these challenges.

Non-random assignment. First, although CareerStart was randomly assigned, the level of CRI, which is an endogenous variable, could not be. If CRI was associated with these outcomes through some other unmeasured correlates of student engagement or achievement, then observed association between CRI and each outcome would be biased or inconsistent and could not be inferred as causal. The Rubin causal model (RCM), also known as the potential outcomes model, provides a causal framework for clarifying the problems of inferring causality from an endogenous variable of interest such as CRI (Holland, 1986; Morgan & Winship, 2007).

Formally, I assumed that the outcome (engagement or achievement) for student i (with I = 1,…, N) under treatment condition j is Y_{ij} with each treatment condition being one

of multiple potential levels of CRI represented by j (of j = 1,…, J; Morgan & Winship, 2007). (For simplicity, I treated j as indicating both the level of CRI and the teacher, and ignored conditions where multiple teachers have the same level of CRI and where a teacher may be conceived of as having multiple levels of CRI.) Second, the student had one potential outcome Y_{ij} under each possible treatment. At most one of these potential outcomes could be realized, as each student could only be assigned to one condition (the fundamental problem of causal inference; Holland, 1986). To circumvent this, the treatment effect was defined as a function of the distributions of students assigned to each possible level of CRI, usually labeled the average treatment effect (ATE). Three assumptions were required. First, manipulability implied that students could be assigned to any condition, which is required for all potential outcomes to be defined. Second, the stable unit treatment value assumption (SUTVA) implied that the potential outcome under any level of CRI was not affected by peers also assigned to that condition. Neither of these could be challenged in the present study and were simply assumed.

The third assumption, unconfoundedness, was directly related to the non-random assignment of CRI. Unconfoundedness implied that each student's assignment (A) to a treatment condition j was independent of his or her potential outcome in that condition (Morgan & Winship, 2007):

$$\text{(i) } Y_{ij} \perp \text{A}$$

If potential outcome Y_{ij}, which describes the student's benefit from instruction infused with various levels of CRI, was correlated with any variable X then condition i fails. A weaker form of this statement conditions on student and teacher background, here represented by X_i and X_j, respectively:

$$\text{(ii')}\ Y_{ij} \perp \mathrm{A} \mid X_i = x_i, X_j = x_j$$

Condition *ii* fails if there were unobserved X_i or X_j that could not be conditioned on when estimating the effect of CRI on student engagement and achievement.

The bias from the non-random assignment could increase or decrease the observed magnitude of the effect and could be related to observed or unobserved variables. For example, more experienced teachers (X_j) may have been more likely to use CRI and could be assigned to higher achieving classrooms of students (as measured by previous performance; X_i). This may have led to higher student engagement or achievement (Y_{ij}) that might be attributed to CRI rather than (more appropriately) to a mix of serendipitous assignment and a greater command of effective teaching strategies owing to experience. This could have magnified the observed effect of CRI relative to what it would be in the population. Alternatively, less experienced teachers may be more amenable to using innovative strategies such as CRI or may be more open to employing theories about human development that were learned in pre-service preparation, but may have been assigned to lower achieving classrooms. These assignments would have led to lower engagement or achievement being attributed to CRI, attenuating the effect.

Despite not being randomized, observational or quasi-experimental approaches could approximate the results of a comparable randomized design if appropriate assumptions were met. In the present study, I propose to use instrumental variable estimation (IVE) to estimate a local average treatment effect (LATE; Angrist, Imbens & Rubin, 1996), which may satisfy the assumption of unconfoundedness, making inferences about CRI causal. I discuss the details of this method when discussing the analysis methods.

Measuring teacher CRI using student perception. Teachers were not asked directly about their use of CRI. Instead, the level of CRI used by teachers was obtained from end-of-year student survey reports of their agreement with the statement that each separate core subject teacher *used career examples frequently*. That is, the actual level of CRI was unobserved, but student perceptions of the level of CRI, which varied across students having the same teacher, was measured. In this design, students acted as raters of their teachers' use of CRI. In previous studies (Orthner et al., 2012; Rose et al., 2012) CRI at the teacher level was obtained by taking the average response of students' perceptions for each teacher.

In this study, a more rigorous measurement strategy has been implemented that accounted for the effect that multiple student, school and home background factors may have on a student's perception of how much CRI is considered "frequent," including students' race/ethnicity, family income, previous achievement, and students' inherent tendency to rate their teachers high or low as measured by previous years' perceptions of CRI. Exploratory analyses demonstrated that students' responses to each question were associated with demographic data such as race/ethnicity and free/reduced price lunch, a rough indicator of family income (Woolley & Rose, unpublished manuscript). In addition, parents' discussions with their children about the work that they do were indicated using five items on the student survey, and theses items were included in the measurement model. This strategy is an innovative approach in that I utilized Empirical Bayes shrinkage to estimate teacher CRI using the teacher residual as the deviation from the grand mean in a multilevel model (Raudenbush & Bryk, 2002).

Data Sources

There were four sources of data. First, student surveys administered at the end of sixth, seventh and eight grades were used to collect a number of important variables including psychosocial information from students describing their emotional and cognitive engagement with school; student reports of their teachers use of career examples; students' reports of career family interactions; teacher rigor and press; and future planning. Second, student academic data, consisting primarily of end-of-grade (EOG) math and reading exam performance from third to eighth grade and demographic data were obtained from school district administrative files. Third, teacher surveys recorded years of experience, race, gender and perspective on career relevance in instruction. Fourth, administrative records contained information only on students while they were within the district. All data collection procedures received IRB approval.

Sample

The sample included a cohort of students beginning sixth grade in school year 2005-06, and their eighth grade teachers.

Student participants. There were 2,866 eighth grade students in 14 eligible schools reporting on each of their core subject teachers: 49% were female; 30% were Black, 15% were Hispanic; 49% were White; 6% were Asian, Aleutian, Pacific Islander or Native American; 52% received free or reduced price lunches; 32% came from single-parent homes; 18% were identified as academically gifted; and 11% received special education services. The average age was 11 years, 6 months.

Teacher participants. The teachers included in this analysis were eighth grade math, language arts, science and social studies teachers for these students. The students were

considered “nested within” each of these four groups of teachers, though with different nesting patterns by subject (for example, the students having the same math teacher may not have had the same language arts teacher). There were 214 eighth grade teachers included in this study, including 62 math teachers, 55 language arts teachers; 46 science teachers and 51 social studies teachers (Table 2.1). Language arts teachers had an average of 15 years of experience; 84% were female; and 18% were Black. Math teachers had an average of 12 years of experience; 76% were female; and 15% were Black. Science teachers had an average of 15 years of experience; 63% were female; and 18% were Black. Social studies teachers had an average of 13 years experience; 56% were female; and 9% were Black. There were no Latino or Latina math teachers and less than five Latino or Latina teachers in any other subject.

Measures

Dependent variables. Several dependent variables of social or developmental importance were examined in this study. School valuing, a measure of emotional engagement to school, was measured using seven items from the Student Identification with School measure (internal consistency reliability $\alpha = .79$; Voelkl, 1996) (sample item: “school is one of the most important things in my life”). The valuing scale indicated students’ belief that school is important to them at present and for their future, including the importance of education in getting a job. Cognitive engagement was measured using the School Success Profile engagement scale ($\alpha = .80$; Bowen, Rose & Bowen, 2005), and consisted of three items indicating that students find school fun and exciting, and that they look forward to going to school and learning new things. Academic achievement was measured using performance on North Carolina end-of-grade reading and math standardized achievement

tests for eighth grade in the 2007-08 school year. Figures 2.1-2.4 are kernel density functions for each dependent variable.

CRI measure and measurement strategy. The treatment or independent variable in this analysis was a measure of the teacher's probability of using of CRI. As previously noted, teachers were not asked directly about their use of CRI. Instead, students were asked the following question for each content area on a survey conducted at the end of each grade level: "In [math/language arts/social studies/science], my teacher often uses examples from jobs and careers." Potential responses to these questions included strongly disagree, disagree, uncertain/not sure, agree, or strongly agree; these were summarized into agree/strongly to indicate frequent and disagree/strongly disagree to indicate infrequent. Uncertain/not sure responses were set to missing and are imputed; this decision is reconsidered in the robustness analysis (Figures 2.5-2.8 are kernel density functions of CRI for each subject). The subsequent measurement strategy: a) accounted for the effect that multiple student, school and home background factors may have had on each student's perception of how much CRI was considered "frequent," and b) summarized this measure to the teacher level. A conditional model included students' race/ethnicity, family income, and elementary school math and reading achievement as predictors of students' perception that the teacher used CRI frequently. Students' inherent tendency to rate their teachers high or low, as measured by previous years' perceptions of CRI, was used to eliminate unmeasured time-invariant influences of perception of CRI. A key set of predictors described the extent to which parents' discuss their own careers with their children during eighth grade; these measures are discussed below.

The multilevel logistic regression model (or hierarchical generalized linear model) estimated each teacher's probability (Woolley & Rose, unpublished draft; Raudenbush & Bryk, 2002). This multilevel model was a regression of student-reported agreement with the CRI question on student and family factors at level 1 (student rater) and a teacher random effect (but no covariates) at level 2 (teacher), with separate regressions for each content area. Because each teacher's use of CRI may include factors that are invariant within the school but should nevertheless be included in the teacher estimate when comparing teachers across schools, school level was excluded from the model such that effects inherent to the school would be absorbed into the teacher measure. Such effects may have included, for example, the principal's enthusiasm for CareerStart or the level of support provided by the district.

The empirical Bayes residual, the deviation of each teacher's mean log-odds of frequently using CRI from the grand mean log-odds of all teachers, was derived as teacher CRI (Raudenbush & Bryk, 2002). The log-odds was converted to the probability that the teacher frequently used CRI and multiplied by 100 to arrive at the *probability of frequently using CRI* used as the independent variable in the analysis models. The CRI effect on the outcome was interpreted as the conditional incremental gain to eighth grade engagement, valuing or achievement from a one percentage point increase in the probability that the teacher used CRI.

Student-level covariates. Student-level characteristics included demographics and family characteristics such as age (and age-squared and cubed), race/ethnicity (Black and Latino/Latina children relative to White children and an unsubstantial number of children of other backgrounds), gender (entered as female), free-reduced lunch, parent marital status (entered as single parent), academic giftedness, and special education. In addition,

achievement on elementary and middle school end-of-grade achievement tests from third to seventh grades, as well as their squares and cubes for addressing nonlinear relationships, were used to control for previous years' learning in the models for academic achievement (math and reading separately). For valuing and engagement, elementary school achievement on math and reading was used alongside sixth and seventh grade reports of the dependent variable as controls.

Five items reporting students' interactions with their parents' regarding their parents' jobs and careers, and one item reporting the students' plans to attend college (*I plan to attend some type of college after high school*), were included in the measurement models for CRI. The questions about students' job and career interactions with their parents included the following questions: a) *my parents/guardian tell me about the kind of work they do*; b) *my parents/ guardian show me the kinds of things they do at work*; c) *my parents/guardian tell me about their jobs*; d) *my parents/guardian have shown me where they work*; and e) *my parents/guardian tell me about things that happen to them at work*. Given the size of the student sample, selection of the variables to use for controls, for which I preferred an exhaustive specification to eliminate all potential autocorrelation and other sources of confoundedness, was driven largely by robustness diagnostics to eliminate collinearity. Although these measures were used in the measurement model, they were not in the outcome models for demonstrating the effect of CRI on engagement and learning, as I assume that these measures may be endogenous. They may be endogenous with respect to CRI and the outcomes as students' perception about how often they hear their parents talk about careers may be conditional on how often they heard their teachers talk about careers, and both may be associated with the outcomes.

Teacher-level covariates. Teacher-level covariates included race/ethnicity, gender, years of service, teacher perspective on career relevance, which were obtained from teacher surveys. In addition, because teachers were not randomly assigned to students, and studies have demonstrated the association between teacher effectiveness and the aggregate previous performance of their students, a crude correlate of teacher assignment was obtained by taking the average seventh grade math or reading score of the students assigned to each eighth grade teacher. To the extent that use of CRI is correlated with teacher effectiveness and student outcomes, this will address an alternative pathway through which CRI could be associated with student outcomes.

The CareerStart logic model suggests an alternative pathway for the effect of CareerStart through teacher effectiveness, an issue I took up in the discussion of the CareerStart logic model. Typically, teacher effectiveness is measured using value-added estimates of student learning, but to do so in this study would have required matches of the eighth grade teachers to their previous years' students, which were not available in the data. Instead, two academic press measures were collected on the student survey, and used in this analysis: *rigor* and *preparation for further schooling*. These were constructed by calculating the sum of students' responses to the following set of questions (each question on a 1 to 5 Likert response scale; yielding a sum ranging from 3 to 15). These questions were originally envisioned as a single construct of academic press (Dennis Orthner, personal communication, December 27, 2012) but for which factor analyses demonstrated a two-factor structure. For rigor the items were (a) my teachers give me challenging homework; (b) my teachers set high standards that I must meet to get good grades; (c) my teachers push me to think harder (α = .72). For preparation for future schooling the items were (d) my teachers are preparing me for

the next grade; (e) my teachers are preparing me to succeed in high school; and (f) my teachers are preparing me for college or other education/training after high school (α = .84). These measures were highly correlated (ρ = .70).

Like CRI, conditional models (consisting of the same student-level measures) were estimated to derive a teacher level estimate of these practices, which was then entered into the analysis models. However, because students were not asked about each specific subject's teacher (item wording indicated "my teachers"), I could not attribute the perceived level of press to any of the four teachers. To get around this measurement problem, the attribution was made empirically, based on the different nesting patterns of students across the four subjects, using the multilevel model. That is, because different groups of students are nested within each subject, there would be variability in how each student's reports of their teachers' press was aggregated to the teacher level. This is a substantial limitation of the current study, and I am cautious about making inferences from this analysis, and instead propose that it raises important measurement issues that should be taken up in a future evaluation of CRI or CareerStart.

School-level measures. School-level covariates included title I status, proportion free/reduced lunch and minority student. The number of school-level covariates that could be entered was severely limited by the small number of schools, and these typically showed the highest amount of residual variance explained in exploratory models and models developed in Woolley et al. (in press).

Missing Data

Missing data occurred throughout the student data, including in the survey questions (from 23 to 1,025 of 2,866 students) and the achievement data. Further, I assumed that

students who responded “not certain/unsure” to the CRI questions on the survey gave effectively missing responses and recoded this response to missing. All of these missing values were subjected to an analysis to determine if they were missing completely at random (MCAR) using the Little (1988) chi-square test, which is a multivariate test of whether students with missing values were different from students without missing values. The hypothesis of MCAR was rejected suggesting that at least some of the data points were conditionally missing.

The proper approach in this context was to subject the missing values to an informed, robust imputation model (Schafer, 1997). This model included, as recommended, all of the analysis variables described previously that were to be used in the analysis and measurement models, as well as a set of auxiliary variables that supplemented the analysis variables in explaining the missing values. Ten imputations of the data were simulated using SAS Proc MI; all analyses were conducted on each imputation separately using SAS and Stata and then the parameter estimates from these distinct analyses combined according to rules developed by Rubin (1976).

Missing teacher level covariates (gender, race/ethnicity, years experience and perspective on CRI) presented an additional challenge. Multilevel techniques for imputation have not been developed, and multivariate normal techniques imputed values for these variables that varied within teachers. Differences between models with and without the teacher covariates could not be attributed to the covariates themselves or to the resulting loss of students and teachers with missing values. Techniques for addressing this problem are provided in the description of robustness tests.

Analysis

First, I conducted the CRI measurement analysis, followed by the outcome analysis. The primary method of analysis used in this study, which addressed the non-random nature of CRI, was instrumental variable estimation (IVE) for obtaining a LATE. Several alternative analyses were also run to examine the robustness of the findings from the IVE to variations in model specification, alternative handling of missing data, and even the assumptions of IVE itself.

Measurement model. As described in the section on the CRI measure, a rigorous measurement model for CRI in each content area was tested and the teacher level CRI measure obtained as the deviation of each teacher from the overall average of all teachers in that content area in the study. The model of P_{sij} = probability (student *i* with teacher *j* in subject s reported CRI was used frequently) took the following form:

$$\text{(1)} \qquad \text{Log}\,(P_{sij}/[1 - P_{sij}]) = \gamma_{00j} + \gamma_{00j}\,\mathrm{X}_{ij} + u_{00j}$$

The index *j* indicates the teacher and *i* indicates the student. The deviation from the overall average is the teacher-specific value of u_{00j}. As needed for diagnostics, the student error variance was represented by the logistic error variance, $\sigma^2 = \frac{\Pi^2}{3}$ = 3.2899 (Snijders & Bosker, 1999); alternatively, (1) was re-estimated as a linear probability model replacing Log(.) with P.

Instrumental variable estimation (IVE). CRI is the teaching practice that CareerStart is intended to promote. Because CRI is non-random it cannot be directly and randomly assigned by the evaluator. With the level of CRI being partly determined by teachers, after assignment to CareerStart, it cannot be guaranteed that there is a 1:1 correspondence between treatment assignment and use of CRI. Econometric or statistical

adjustments are needed to obtain a causal estimate from a non-randomly assigned treatment, yielding a type of treatment on treated estimate known as a local average treatment effect (LATE) rather than an intent-to-treat population average treatment effect (ATE). One method used to obtain a local average treatment effect is IVE (Angrist, Imbens & Rubin, 1996).

A LATE is, intuitively, a ratio consisting of the effect of CareerStart on the ultimate outcome (e.g., math achievement) divided by the effect of CareerStart on promoting teachers' use of CRI. This ratio is also known as the Bloom (1984) treatment-on-treated estimator. If the "dose" of CRI was 100% the level promoted by CareerStart, this quotient would have yielded a CRI effect equivalent to that of the CareerStart intent-to-treat effect. The denominator being less than 1 in effect inflates the observed or intent-to-treat effect of CareerStart on the outcome. The resulting ratio could be interpreted one of two ways. First, it could be interpreted as the effect of teacher compliance with assignment to treatment on student outcomes. Or it could be interpreted as the effect of a "dose" of CRI on the student outcomes. Which interpretation is chosen depends on whether one views teachers as the receivers or implementers of the treatment in the context of the purpose of the analysis. As a mediator in the logic model, CRI is both, and in the present context, the purpose was two-fold: to validate the logic model, which supported the teacher as receiver (of CareerStart) interpretation; and to contribute to the scholarship on relevance, which supported the teacher as implementer (of CRI, and more generally, relevance) interpretation, meaning that both interpretations had merit.

An IVE estimand could be calculated directly via this ratio, but it would have been an unconditional estimand, unadjusted for measured covariates of the treatment effect. To adjust for measured non-random factors, the IVE estimand could be obtained via a two-stage

regression method; interpretations about compliance or dosage effects remain valid. To meet the conditions required for IVE, a variable (referred to as an instrument) had to be identified that was: highly correlated with the endogenous variable, but not independently associated with the outcome. This second requirement, known as the exclusion restriction, could equivalently be stated as requiring that the instrument had no association with the error in the model of Y on CRI; and that all of the effect of the instrument on the outcome was through the endogenous variable (a weaker form allows for conditioning on measured covariates). The exclusion restriction in effect describes the process by which the instrument causes the outcome.

The literature on LATE estimation using IVE suggests the randomized treatment assignment variable (CareerStart) as the instrument (Angrist, Imbens & Rubin, 1996). In fact, CareerStart satisfied both IVE requirements: it was associated with CRI (Rose et al., 2012) and, being randomly assigned, had no association with the error between CRI and the outcome. To rely on the random assignment as the instrument, Angrist, Imbens and Rubin (1996) showed that additional assumptions are needed. SUTVA, an assumption of the RCM, had to be assumed to be satisfied. Further, the treatment had to have a unidirectional effect on participants' CRI potential outcome; that is, CareerStart could only promote higher CRI or it could only promote lower CRI, it could not be assumed to do both. In other words, all participants' potential outcome under CareerStart had to be a frequency of CRI use that was higher than their potential outcomes under the control condition. This is the monotonicity assumption, and because it regards potential outcomes, it could not be observed or proven from observed data (Angrist, Imbens & Rubin, 1996). However, Rose et al. (2012)

demonstrated that the average observed teacher response to CareerStart was to more frequently use CRI, providing some evidence that the assumption is realistic.

In this study, I used panel multilevel IVE methods in a two-level clustered multilevel design of students nested in teachers. Nesting of teachers within schools was not accounted for. Because IV estimation also inflated the standard error of the focal variable much like multilevel modeling does, it is not clear the extent to which the standard error was upwardly or downwardly biased.

The two-stage model for CRI on each outcome Y is as follows. First, a linear probability model predicted $P(CRI_{sij})$ with r_{ij} as student error and ξ_j as teacher error at this stage:

$$(2) \qquad P(CRI_{sij}) = \alpha_{s00} + X_{ij}\alpha_{sxi} + X_j\alpha_{sxj} + \alpha_{CS}CS_j + r_{ij} + \xi_j$$

The predicted value of $P(CRI_{sij}), P(\widehat{CRI}_{sij})$was then entered into a second regression model as the independent variable in place of its observed equivalent, with e_{ij} and v_j as student and teacher errors, respectively:

$$(3) \qquad Y_{ij} = \pi_{00} + X_{ij}\pi_{xi} + X_j\pi_{xj} + P(\widehat{CRI}_{sij})\pi_{CRIj} + e_{ij} + v_j$$

The coefficient tested for the effect of CRI on each outcome and for answering questions 1 (does CareerStart promote higher engagement and valuing) and 2 (does CareerStart promote higher end-of-grade math and reading achievement) was π_{CRIj}. In alternative models for answering question 3 (does teacher press represent an alternative pathway for the effects of CareerStart), rigor and preparation for further schooling were each separately entered as the endogenous variable, and subjected to the same two-stage model as represented by equations 2 and 3. The coefficients π_{Rj} and π_{PFSj} were obtained from these models and similarly tested.

Two models were tested. In the first all student-level predictors and teacher-level predictors that were derived from students, including the teacher assignment proxy and the CRI measure (or alternatively, rigor or preparation for future schooling), and a small set of school level measures such as average FRL or race/ethnicity composition. In the second model teacher's years of experience, gender and race/ethnicity were all added to the model. Both of these models are given equivalent weight in the interpretation of the findings; the former suffers from an incomplete specification as these teacher variables are associated with use of CRI and thus potentially confounders, and the latter model suffers from unaccounted for missing data on these important measures. A robustness check was used to ascertain whether the differences between them were due to confoundedness or the loss of records with missing data. Two stage models were estimated with the XTIVREG procedure in Stata 12 (Statacorp, 2011).

Robustness Analyses

A number of subjective decisions, such as how to handle students' responses of *uncertain/not sure* to the CRI question and how to handle missing teacher covariates, were employed in these analyses and thus the study design was subjected to a variety of robustness tests to examine the sensitivity of the findings to these decisions. The imputation of *uncertain/not sure* responses after re-setting them to missing was reconsidered. Alternatively, *uncertain/not sure* was included as part of not agreeing to the statement that students' teachers frequently used career relevance as a form of passive non-agreement (this is how the response was handled in Rose et al., 2012). The missing teacher covariates, which presented a complex problem due to the unavailability of multilevel imputation techniques that would impute invariant-within-teacher values of the missing covariates, made it difficult to discern

if effects observed after entering these covariates were due to the covariates themselves or due to the loss of records with missing values. This problem was subjected to two robustness tests. One was to impute the covariates as if multivariate normal and average the resulting imputed values up to the teacher level. The second was to remove the records with missing teacher covariates from the analysis and run the analysis without the teacher covariates.

Multiple versions of the models under different student covariate specifications were tested, without substantive changes to the findings, though completely unconditional models showed some differences. A final robustness analysis was run under the assumption that conditional on the model covariates, CRI was not endogenous. For the purpose of most accurately reflecting the multilevel structure of the data these robustness models included a school level in addition to student and teacher levels. The addition of this level, which added a school-level error to the model, should have reduced the teacher-level conditional variance, potentially making the estimation of CRI, a teacher level variable, more precise. Without IVE, the standard errors of CRI should be further clarified.

Results

First I describe the results of the CRI measurement model, then turn to the results of the outcome analysis, including the IVE and the robustness checks.

CRI Measurement Results

The variance in the student perceived CRI was decomposed into student and teacher components to assess the extent to which there was actual teacher variation in use of CRI, and how much of the variance was related to students' perceptions (Table 2.2). The proportion of variance at the teacher level was 0.08 for language arts and science CRI, 0.11 for math, and 0.06 for social studies. For robustness to the assumption of a logistic

distribution, a linear probability model, which does contain a residual at the student level was estimated and the decomposition showed that the proportion of variance at the teacher level was 0.05 for language arts and science, 0.07 for math, and 0.03 for social studies. These results indicated that 89-90% or more of the variance was between students, suggesting the need to control for this variation to estimate a measure of eighth grade teacher CRI.

Conditional models containing all available student predictors showed that, students' sixth and seventh grade reports of CRI, students' reports of how often they heard their parents talk about their careers at home, and students college plans were the most robust predictors of students' perceptions, eliminating nearly all unconditional associations between CRI and demographics (Tables 2.3-2.6). With a discrete measurement model of high vs. low use of CRI, I report these results as pertaining to those characteristics that are significant in increasing the probability of reporting high use of CRI or the probability of reporting low use of CRI, in each subject. Across all CRI measures, student reported sixth and seventh grade CRI was significant at .01 or .001, with those students reporting high CRI in these grades having 45-180% higher odds of reporting high CRI in eighth grade.

Students reporting that they agreed that their parents/guardians showed them the kinds of things they did at work (OR = 1.58; $p < .01$) and students reporting that they planned to attend college after high school (OR = 2.79; $p < .001$) reported high language arts CRI (Table 2.3). Students reporting that their parents/guardian told them about the kind of work that they do (OR = 1.90; $p < .05$), or showed them the kinds of things they did at work (OR = 1.44; $p < .05$), and students planning to attend college (OR = 2.58; $p < .001$) reported high math CRI (Table 2.4). However, Latino or Latina students also reported high CRI (OR = 1.49; $p < .05$). Students reporting that their parents/guardians showed them the kinds of

things the did at work (OR = 1.68; $p < .05$) or told them about things that happened to them at work (OR = 1.68; $p < .05$) and students reporting that they would attend college some day (OR = 2.32; $p < .01$) reported high science CRI (Table 2.5). Students reporting that their parents/guardians showed them the kinds of things they did at work (OR = 1.50; $p < .05$) and students planning to attend college (OR = 2.49; $p < .01$) reported high social studies CRI (Table 2.6). However, students receiving free or reduced price school meals also reported high social studies CRI (OR = 1.37; $p < .05$); further, a one point increase in fourth grade reading EOG predicted a 2% lower odds of reporting high social studies CRI. These results suggest the importance of including reports of CRI in prior years (to control for time-invariant influences of students' perceptions), elementary school math and reading achievement, and students' reports of plans for future schooling and discussions with their parents/guardians about jobs and careers. In contrasting these findings with exploratory analyses, these groups of variables remove most, though not all, of the associations that were previously observed for race/ethnicity, family income and gender.

Analysis Results

Four variations on CRI models are presented here. These include IVE with no teacher covariates, IVE with teacher covariates, multilevel models without IVE, and models that treated uncertain responses as passive disagreement at the imputation stage. After discussing the main findings and robustness tests to answer the questions about CRI, alternative pathways for CareerStart, through rigor and preparation for future schooling, are explored, to answer the third.

IVE, no teacher covariates. In the two-stage IVE models for estimating LATE, using the first imputation in which teacher covariates were not imputed but uncertain

responses to the CRI question were imputed, CRI did not demonstrate a significant and promotive effect on math, reading or engagement. However, findings for valuing were significant and robust across multiple forms of CRI in models in which teacher covariates were not included (Table 2.7A). A .01 increase in the probability that the language arts (Table 2.19) or the science teacher (Table 2.20) used CRI yielded a .01 increase in student valuing ($p < .001$); a one percentage point increase in the probability that the social studies teacher (Table 2.21) used CRI also increased valuing by .01, but with less precision ($p < .01$). These findings are presented as the main findings regarding the effect of CRI in promoting valuing, based on the findings of robustness tests that follow.

IVE, teacher covariates. The first test of the robustness of these findings was to add teacher covariates to the models. When teacher covariates were added, only social studies (Table 2.7A and 2.21) CRI promoted higher valuing ($\beta = .01$; $p < .05$). However, the diminution in significant findings could have been due to the loss of teachers having missing values, as these values were not imputed, rather than due to the effects of the covariates. Two approaches confirmed that the change was at least partly due to the loss of data from students having teachers with missing data. First, the model in which teacher variables were imputed at the student level (the second version of the imputation) and averaged up to the teacher (Table 2.7B) showed that all forms of CRI were predictive of higher valuing, including math ($\beta = .01$, $p < .05$); language arts ($\beta = .01$; $p < .001$); science ($\beta = .01$; $p < .001$); and social studies ($\beta = .01$; $p < .01$). This would not have been the case if the change had been due to the confoundedness.

Second, the model without teacher covariates, but also without the students having teachers with missing values showed that both math and language arts ($\beta = .01$; $p < .05$) were

significant for valuing (see Table 2.7C). The math finding was not found in the original model without the students having missing teacher covariates deleted; and science and social studies findings were found. This provides more evidence that it was deletion of missing values that drove most of the change in findings from models without teacher covariates to models with teacher covariates. Only language arts was unchanged, providing the only source of evidence that the teacher covariates may have confounded the treatment effect. On the whole, the addition of teacher covariates did not detract from characterizing the previous findings as the main findings for CRI.

Models without IVE. Another concern of these models was that using IVE, which could have inflated standard errors, may have been more likely to result in type II errors. A robustness check confirmed that the IVE may have controlled for important unobserved variation and that any inflation in standard errors did not increase type II error. Under the assumption that the full conditional model explained all of the confoundedness from non-random assignment of CRI, multilevel models without instrumental variables were estimated for each CRI-outcome pair. In one version of each model, teacher covariates were included, while in another they were left out. A standout finding was that math was significant ($\beta = .04$; $p < .05$) in the model without teacher covariates (Table 2.7D). Valuing again emerged as a robust finding. Math CRI in models with ($\beta = .01$; $p < .05$) and without ($\beta = .01$; $p < .05$) teacher covariates predicted higher valuing; language arts CRI without teacher covariates ($\beta = .01$; $p < .05$); science CRI in models with ($\beta = .01$; $p < .05$) and without ($\beta = .01$; $p < .05$) teacher covariates; and social studies CRI without teacher covariates ($\beta = .01$; $p < .001$). It was noteworthy that the magnitudes and directions of the IVE coefficients and the coefficients from the multilevel models were very similar. However, they were not similar

enough to reject that the IVE added additional control for unobservables. I take this up further in the discussion.

Uncertain as passive disagreement. In the third version of the imputation in which students' responses of "uncertain/not sure" to the CRI question were assumed to constitute passive disaffirmation that their teachers frequently used career examples, valuing results were observed for models with and without teacher covariates (Table 2.7E). Math CRI was a significant predictor of valuing ($\beta = .01$; $p < .05$) with teacher covariates; language arts CRI was a significant predictor of valuing ($\beta = .004$; $p < .05$) without teacher covariates; science CRI was a significant predictor of valuing with ($\beta = .005$; $p < .05$) and without ($\beta = .004$; $p < .01$) teacher covariates; and social studies CRI was a significant predictor of valuing with ($\beta = .005$; $p < .05$) and without ($\beta = .01$; $p < .01$) teacher covariates. These findings suggest that the main model results are robust to the approach used to handle students' responses of *uncertain/not sure* on the CRI questions.

Academic press. The logic model for CareerStart indicates that teacher effectiveness, measured in this study using two dimensions of academic press (rigor and preparation for future schooling), may be an alternative process by which CareerStart promotes engagement and learning (Table 2.22). Before presenting the findings, the coefficients were scaled down to the 0-1 metric that CRI was on, to make the coefficients comparable. Using the IVE framework (and treating each press measure separately as the endogenous variable), higher valuing was predicted by rigor with similar effects for math ($\beta = .024$; $p < .05$), language arts ($\beta = .032$; $p < .001$), science ($\beta = .031$; $p < .001$), and social studies ($\beta = .029$; $p < .001$). Higher valuing was also predicted by preparation for future schooling in math ($\beta = .042$; $p < .001$), language arts ($\beta = .036$; $p < .001$), science ($\beta = .040$; $p < .001$), and social studies ($\beta =$

.034; $p < .001$). Preparation for future schooling also predicted higher engagement including in math ($\beta = .029$; $p < .05$); science ($\beta = .031$; $p < .05$); and social studies ($\beta = .016$; $p < .05$). The findings for each subject were similar but not identical. The similarity was because students reported a single perception measure across all teachers and I derived a subject-specific teacher estimate by aggregating the perception to the teacher level using the different nesting patterns that students demonstrated across teachers.

Discussion

In this study, I examined the causal effect of relevance as a teaching practice, in the form of career examples associated with lesson content in the four main subject areas of math, language arts, science and social studies, on students' engagement, valuing and end-of-grade achievement. I determined whether CRI serves as the process by which a randomly-assigned teaching practice intervention, CareerStart, promotes student engagement (Orthner et al., 2012) and math achievement (Woolley et al., in press). A challenge in the effort to ascribe causality to career relevance is that the practice itself could not be randomly assigned. Teachers who were randomly assigned to each CareerStart treatment condition could choose to ignore their treatment assignment. Some teachers in schools assigned to CareerStart, for example, elected not to use career examples in their lessons or give incomplete lessons (Woolley, Rose, Mercado & Orthner, 2013). The challenge to evaluation arises because if the choice to ignore assignment was non-random and associated with the treatment outcomes, this could provide an alternative explanation for the observed effect of CRI. A non-experimental evaluation method, IVE, combined with the randomized assignment of CareerStart, was used to estimate a local average treatment effect (LATE). The LATE could be interpreted as the causal impact of CRI, provided that assumptions are met regarding the

association between CareerStart and CRI (Rose et al., 2012) and that the assignment of CareerStart was random. Combined with a strong theory of change for CareerStart and several robustness checks, the evidence from this analysis lends support to the causal interpretation of CRI on student valuing. However, some of the results raise important questions that can only be answered by future studies.

Questions 1 and 2 concerned whether CRI promoted higher engagement and valuing, and whether CRI promoted higher end-of-grade math and reading achievement, respectively. Valuing, adapted from the Student Identification with School measure (Voelkl, 1996) represents a measure of emotional engagement to schooling. Students who report higher valuing believe that schooling will help them achieve their objectives and that this belief manifests affectively (Voelkl, 1996). Valuing was the most robust finding in this analysis, with CRI showing significant and positive effects for valuing across multiple subject areas under a variety of specifications and robustness checks, and the effects were all of similar size, at about .01. Social studies CRI was a significant and positive predictor of valuing in all specifications and robustness checks except for the model without IVE having teacher covariates. Language arts CRI was significant except in models that included non-imputed teacher covariates, though the evidence of robustness checks indicated that the loss of missing values may explain the differences between models with and without teacher covariates, rather than the confoundedness of the teacher covariates. Math CRI was a significant predictor of valuing only in robustness checks that included imputation of teacher covariates, models without IVE, and in both cases, only when uncertain responses to the CRI question were imputed rather than considered passive disagreement. Science CRI was significant and positive in a model without teacher covariates, and in several robustness

checks, including those with imputed teacher variables, no IVE, and models in which uncertain responses were considered passive disagreement.

The magnitude of the effect of CRI on valuing can be illustrated by calculating the difference in valuing scores for two students having teachers at one standard deviation above and below the mean in the probability of using CRI. For language arts, the gap between these two teachers was 18.2 percentage points, which translates to a .182 difference between these two students on the valuing scale, which ranged from 1 to 5 with a standard deviation of .64, suggesting that the difference between having teachers at -1 and +1 standard deviations amounted to about one-quarter of a standard deviation in the outcome. For math, the gap was 22 percentage points (.22 difference in valuing); for science 17.4 percentage points (.174 difference in valuing); and for social studies 17 percentage points (.17 difference in valuing).

These findings are supported by other investigations. Orthner et al. (2012) examined the effect of CareerStart on valuing and found that being in a school in the treatment condition increased student valuing by 1.41. The evidence from this same study suggested an association between CRI—measured by the unconditional teacher mean of the original Likert responses to the question about the frequency of use of career examples—and valuing was 1.31. This indicated that a one-point change on the CRI scale (ranging from 0 to 4) would result in a 1.31 point change in valuing.

The findings described above were robust to specifications regarding imputation, teacher covariates, and instrumental variable modeling. Two other findings for CRI emerged that, although not robust across multiple specifications, suggest that further investigations are required. CRI predicted higher math achievement in the version of the model without IVE. The IVE model showed an effect of .031, and the model without IVE, .039. This was the

entirety of the difference between the two findings, as the standard error was the same (.018) in both models, contrary to what was expected of IVE (i.e., inflating standard errors). The t test in the IVE model suggested a trend ($t = 1.67$; $p < .10$) but not a finding at the desired level of confidence.

Not finding a math result at $p < .05$ raises questions about how CareerStart may have affected math achievement (Woolley et al., in press), as that is part of the objective of this study. Rose et al. (2012) showed that CareerStart also promoted higher CRI in math. Taken together, the findings of the two preceding studies partially confirmed the logic model and suggested this study as a direction for research. The trend-level finding does not disconfirm that CRI may have been a process for math achievement, and suggests that this issue may need to be taken up again in a future evaluation, but at present the evidence is inconclusive, and therefore other factors must be considered. It may be that the effects of CareerStart observed in previous studies on math achievement were driven by the effect in an earlier year of the treatment, such as sixth or seventh grade. Because student performance in these years is controlled for in the present study, the eighth grade teachers may not have added further to the gain from these years. In addition to being cumulative over multiple years, math performance may also be a function of cumulative CRI across subjects. Due to limitations in the data (students could not be matched to their teachers in sixth and seventh grades), neither of these was examined in the present study; both could be considered in further studies.

Finally, engagement was not significantly affected by CRI except in robustness checks where teacher level imputation (science CRI) or non-IVE models (social studies CRI) were used. This finding may be congruent with the results of previous research into the relationship between CareerStart and engagement, and CRI and engagement. In Orthner et al.

(2012), CareerStart was shown to not predict higher engagement. However, CRI, measured using the unconditional teacher mean of students' responses to the CRI question, was significant, with a coefficient of 1.59 (one unit movement on the 5-point CRI scale translated to a 1.59 point increase in engagement). Orthner et al. (2012) did not use a conditional measurement strategy to estimate teacher CRI. Furthermore, the authors estimated a "teacher team" CRI, combining the responses that students gave to all four subject area teachers, rather than attempting to isolate the effects of individual subjects as in the present study. Finally, Orthner et al. (2012) did not claim to estimate a causal effect, and the authors did not use an econometric method to adjust for unobserved confoundedness.

It is conceivable that the difference between the findings in this study and the associations demonstrated by Orthner and colleagues may in fact be the result of unmeasured confounders in their models. The results of the analysis for answering Question 3, which suggests an alternative pathway through preparation for future schooling, may provide some insight into this relationship. Using IVE techniques, and treating preparation for future schooling as an alternative process for CareerStart, preparation for future schooling was demonstrated to have a significant and promotive effect on engagement. Engagement was measured by a scale consisting of items indicating a cognitive investment of resources in schooling. Positive responses to these questions suggest that students understand the value of school work, something which may be immediately responsive to teaching strategies that link the coursework being learned currently to schooling that will be required at a later time. Results for engagement and preparation for future schooling are robust across most subject areas (language arts, science and social studies). Inferences about this finding should not be overstated, because there were problems with measurement of preparation for future

schooling (in particular, that students were not asked to report on each individual teacher). However, they do suggest a direction for future research into teacher strategies that promote cognitive engagement.

Similarly, valuing was shown to respond, robustly, to both preparation for future schooling and rigor. This finding suggests that multiple teaching practices promote valuing and the IVE suggests that CareerStart's effect may operate through these multiple pathways. Although the LATE IVE method cannot adjudicate between the three the relative magnitudes of the three effects can be examined in a less formal manner. When CRI, rigor and preparation for future schooling were entered simultaneously into non-IVE multilevel models, preparation for future schooling was significant when paired with math, science and social studies CRI; none of these forms of CRI were significant. This finding raises questions about whether CRI was the process by which CareerStart operated to promote higher valuing. Because of measurement problems, it cannot be interpreted as a final answer.

Limitations

There are several limitations in the present study. The first several are measurement limitations. First, although the logic model for CareerStart, informed by the literature on student motivation, indicates that relevance operates through motivational processes, this component of the logic model has not yet been tested. To properly account for the mechanism influencing change in student outcomes, these motivational processes, including expectancy value, goals, interests, extrinsic and intrinsic motivation, and student identity, should have been directly measured. Future evaluations of CareerStart will address this discrepancy. Second and related, motivational processes occur in the context of an underlying developmental framework (Wentzel & Wigfield, 2007), and a panel design

consisting of multiple measures of CRI and each outcome over the full set of available years (e.g., third through eighth grades), would have helped to isolate the treatment effect from these processes. However, although student exam scores were available back to third grade, engagement and valuing were available only as early as sixth grade. Further, although CRI was collected from students as early as sixth grade, the information needed to match students to their teachers was only available in eighth grade. Consequently, I could not replicate the longitudinal analyses for the CareerStart effect in Orthner et al. (2012) and Woolley et al. (in press), and used a model consisting of eighth grade outcomes only. To address this limitation, I used all available covariates at the student and teacher level, and for continuous variables I included both quadratic and cubic terms. A third concern had to do with measurement of academic press in the form of rigor and preparation for future schooling. Students were asked to report on their teachers in general rather than on each specific teacher. In future studies of CareerStart or similar interventions or processes such as CRI, additional measures of teacher effectiveness, including academic press, should be collected and students should be asked to report on each subject's teacher separately. There were limitations in how CRI was measured. For example, students were asked at the end of the year to recall whether teachers used career examples at any time during the school year. More timely measures at multiple time points may have provided better information about teachers' use of career examples. It is also widely understood that EOG scores are imperfect measures of learning, and caution is warranted when interpreting these findings.

A final set of limitations pertains to the IVE method. Some scholars have questioned the view that random assignment satisfies the assumption of IVE that the instrument (CareerStart assignment) not be associated with the outcome through unmeasured processes.

For example, Heckman (1997) indicated that because study participants make assignment-conditional decisions for non-random reasons that are not measured, this assumption may not be satisfied. Further, because the monotonicity assumption concerns potential and not observed outcomes, it is not possible to demonstrate that the assumption holds in the data. Finally, although rigor, preparation for future schooling, and CRI were all shown to be causal processes for valuing using the IVE, and although the relative strengths of each process were demonstrated in a separate multilevel model without IVE, a true multiple endogenous regressor model would have included multiple instruments (e.g., Gennetian, Magnuson & Morris, 2008). The conditions for a multiple instrument model were not met in the current study.

Implications for Further Research

This study contributes in three ways to education policy research into teaching practice. First, it suggests the importance of collecting direct measures of teaching practices. Second, it demonstrates the utility and limitations of using instrumental variables to estimate causal processes in intervention research in schools. Third, it supports the continued investigation of relevance as a teaching practice.

Collecting Practice Data. The ability to draw conclusions from this investigation was limited by the unavailability of all but a handful of indicators of practice. Non-experimental education policy studies typically used structural inputs of teacher preparation and examine these inputs for their association with student learning or teacher effectiveness. These characteristics include board certification (Goldhaber & Brewer, 2000; Hammond, Berry & Thoreson, 2001), the number of content area courses (Boyd, Grossman, Lankford, Loeb & Wyckoff, 2009), PRAXIS II exam scores (Xu, Hannaway & Taylor, 2011), college

entrance exam scores (Harris & Sass, 2011), route of preparation to teaching, such as through traditional preparation (Henry et al., 2011) or lateral entry such as Teach for America (Xu, Hannaway & Taylor, 2011; Glazerman, Mayer & Decker, 2006), or the number of in-service hours of training (Harris & Sass, 2011).

All of these inputs to learning must ultimately operate through the process of teaching. Studies using surveys, teaching logs or observations of both teacher training and teacher practice confirm our intuition that when they are rigorously measured, instructional practices are more robustly associated with achievement than are pre-service training and qualifications (Hamre et al., 2012; Kahle, Meece & Scantlebury, 2000; Meece, 2003; Palardy & Rumberger, 2008). In this regard, the present study, in which relevance-oriented practice was measured strictly via students' perceptions taken at the end of the school year, does not go far enough. Multiple measures of career relevance, including teaching logs, teacher self reports, observations and student feedback at multiple time points during the school year could be developed for future studies of CareerStart or similar practice interventions.

Using IVE to examine teaching processes. Instrumental variable methods are well-known and used frequently in econometric studies, but have not traditionally been employed in psychology, education or social work. However, this is beginning to change with scholars promoting instrumental variable methods as an observational evaluation method appropriate across a wide range of domains (Rose & Stone, 2011; Schochet, 2011), including developmental and psychological research (Gennetian, Magnuson & Morris, 2008). Because IVE methods use standard correlational approaches, the most burdensome aspect of these methods typically has to do with justifying that the chosen instrumental variable satisfies a set of onerous identification conditions. Specifically, the instrument must simultaneously be

predictive of the process and uncorrelated with error. The second condition can be equivalently stated that the instrument must not be associated with the dependent variable except through the specified process, though this can be conditional on other measured variables. The difficulty in establishing that these conditions apply may explain the lack of widespread adoption of IVE outside of econometrics (e.g., Reiss, 2008).

A subset of IVE that can be applied within randomized experiments to examine the causal influences of processes of intervention programs shifts the identification challenge from the instrument to randomized assignment and the Rubin causal model (Angrist, Imbens, &Rubin, 1996). If the random treatment assignment can be shown to be predictive of only higher values of the process and if SUTVA is satisfied, then the assumptions of the IV are satisfied. This study demonstrated the practicality and simplicity of their use in studying teaching practice to establish causality when teachers do not comply with assignment, though there were unresolved challenges when it came to the possibility of multiple potentially non-additive processes.

Relevance as a teaching practice. Research on motivation suggest that making lesson content and tasks relevant to students' current and future lives may act on known motivational processes to promote engagement and student learning. Although motivation was not measured in this study, relevance was measured and shown to be predictive of student valuing, a measure of emotional engagement to schooling. This supports the results of other experiments into relevance oriented teaching strategies. Cordova and Lepper (1996) in a pivotal study on relevance demonstrated that the more they embellished the learning content and tasks with fantasy concepts that were subjectively relevant to the learner, the more the students learned relative to a control condition with no fantasy context. Future

relevance, such as the career relevance examined in this study may have greater value-added to student learning than other forms, because the form of relevance is partly tied to the purpose of schooling. Students now make curriculum choices as early as middle school—choices that may constrain the careers available to them as adults—and are increasingly being exposed to college preparation at this age (e.g., Fleming, 2011). North Carolina Governor Pat McCrory has called for increased vocational training in public schools and to ensure children make informed choices, giving them information in middle school may be helpful to selecting course electives. Career relevance may inform students' identities and goal aspirations in addition to their interests, utility value and intrinsic motivation (Akos & Galassi, 2004; Akos, Konold & Miles, 2004). Husman and Lens (1999) and Husman, Derryberry, Crowson and Lomax (2004) demonstrate that if the future use for a learned task can be imbued with value for the present by associating the learning of the task with future success, it can promote intrinsic interest. Identity formation may also be central to students' motivation, as students who could imagine themselves as successful adults and tie these images to present day motivation to learn, demonstrated better engagement among other outcomes (Oyserman, Terry & Bybee, 2002).

In the context of these contributions to understanding the role of relevance as a teaching practice, the results of this investigation undoubtedly provide some answers. However, they also suggest new questions that can be answered in future studies on the current sample or in a new RCT. First, it is not clear what subjective motivational processes were operating to transmit the benefits of career relevance to valuing. Second, it is not clear what component of the intervention compelled teachers to change their practices; for example, it could have been new knowledge about the motivational value of relevance, or it

could have been changed beliefs and attitudes towards existing knowledge about motivation. Third, an important part of the intervention was to encourage teachers to develop their own career relevant lessons, and because information about such activity was not collected, the influence of such developments on the efficacy of CareerStart is not known. Fourth, as I have demonstrated that rigor and preparation for future schooling, two measures of academic press, promote higher valuing, and preparation for future schooling promotes higher engagement, the CareerStart logic model may need to be clarified to reflect this new knowledge.

Implications for Education Policy

Teachers are now widely considered to be the most important input to learning, and they are assumed to have the potential to compensate for the deficits that students from disadvantaged backgrounds face (Nye, Konstantopolous and Hedges, 2004). Using data from the Tennsessee STAR experiment and a variance decomposition method, Nye, Konstantopolous and Hedges (2004) found that teacher effects for students from kindergarten through third grade range from .12 to .135 for math achievement and .066 to .074 for reading achievement, which constituted about 1/3 a standard deviation in math and ½ a standard deviation in reading between teachers at the 25th and 75th percentiles. Rockoff (2004) used teacher fixed effects models to show that teachers promoted higher student achievement in reading vocabulary (teacher fixed effect joint significance F = 4.43); reading comprehension (F = 2.74); math computation (F = 3.72), and math concepts (F = 5.30). Gordon, Kain and Staiger (2006) found a 10-percentile difference in achievement between students having teachers in the top and bottom quartiles of the distribution of teacher effectiveness. This

suggests, among other policy instruments to be examined by education policy scholars, a focus on identifying and disseminating effective practices.

The development and implementation of evidence-based best practices is a goal of many professions including medicine (Sackett et al., 1996), social work (Albers & Kratchowill, 2006), and public policy (Davis & Boruch, 2001). Advocates for evidence-based education practice call for greater empiricism (Davies, 1999; Oakley, 2002). The subject is controversial, however, with some scholars suggesting it narrows the range of what constitutes acceptable knowledge in education policy (Hleboswith, 2012; Stevenson, 2011; Webster, 2009) with some suggesting it excludes teaching practice altogether (Elliott, 2001). However, any failure to closely examine teaching practice is a sign of myopic application of the principles of best practice rather than of best practice itself. A requirement that educational policies and practices be evidence-based is the impetus for the What Works Clearinghouse, an arm of the Institute of Education Sciences that collectivizes scientific knowledge about effective education policies, including in-classroom practices that promote teacher effectiveness and student learning.

Teaching is a largely uniform activity informed by cultural norms, but produces greatly different results according to context (Sykes, Bird & Kennedy, 2010). Evidence-based practices, applied to teaching, provide scientific support rather than authoritarian or cultural rules for the complex preparation and implementation demands of teaching practice that influence student learning. Historical evidence from the “process-product” literature suggests that teaching practices influence learning, and many of these studies identified effective practices (see Brophy & Good, 1986), though many of these studies were criticized for being insufficiently rigorous and relying largely on correlational designs (Hill, Rowan &

Ball, 2005). Better evidence comes from rigorous designs such as experiments or the use of rigorous non-experimental methods (Cohen, Raudenbush & Ball, 2003; Rowan, Correnti & Miller, 2002), and designs that reflect the shifting goals, priorities and conditions in the classroom environment (Clements, 2007), and the greater intensity and scope of teachers' responsibilities (Valli & Buese, 2007). In addition to strict attention to rigor in the methods used, studies that inform best teaching practice build on the evidence demonstrating the processes by which children learn (Hill et al., 2008), including the objective processes of engagement (Fredricks, Blumenfeld & Paris, 2004) and its subjective precursor, motivation (Wentzel & Wigfield, 2007).

Studies into student motivation, engagement and developmental patterns suggest that interventions that aim to improve engagement in school and prevent school dropout should be implemented in a developmentally appropriate manner at a time before disengagement patterns set in, suggesting that middle school may be an appropriate setting (Janosz, Archambault, Moritz & Pagani, 2008; Roeser, Eccles & Sameroff, 2000). Programs that target student motivation through values, interests and identities have the potential to take advantage of students' natural motivational predispositions to intrinsic rewards such as self-fulfillment and autonomy (Hidi & Harackewicz, 2000). Intrinsic motivation can be sustained, and thus has greater potential for long-term influence on student learning than extrinsic rewards or sanctions, which are typically not sustained (Ryan & Deci, 2000) and may reduce the efficacy of intrinsic reward systems (Hidi & Renninger, 2006). One potential avenue for promoting relevance is to reverse the disconnection between learned content and its potential uses, which typifies how instruction is delivered, by making explicit to students the relevance

of what they are being asked to learn (Cordova & Lepper, 1996; Husman, Derryberry, Crowson & Lomax, 2004).

This study suggests that CareerStart may be one such program. Experimental (Orthner et al., 2012; Rose et al., 2012; Woolley et al., in press) and non-experimental evidence (in this study) suggest that a relevance program implemented in middle school may promote school valuing. As revealed by this study, there are questions about the teaching practices that are ultimately responsible for the observed effects, though relevance has not been completely ruled out, and it is not known which motivational processes are largely responsible for the influence on valuing, or whether altering these patterns prevents dropout from middle school (as students in this cohort have not reached graduation age yet). However, additional evidence does suggest that high school outcomes are improved (Woolley, Unick, Rose & Orthner, under review). This study focused on the individual contribution of CRI to outcomes that could be explicitly linked to that form of CRI (*e.g.*, math CRI on math achievement, or any form of CRI individual on engagement). This addresses concerns that learning in different content areas may reflect different processes (*e.g.,* Hill, Rowan & Ball, 2005; Shulman, 1987). Orthner et al. (2012) combine all four content areas into a single omnibus measure of CRI and showed that it predicted higher engagement and valuing. As noted, this CRI measure was an unconditional average at the teacher team level, and econometric methods were not used to rule out unmeasured confounders, and in particular the influence of academic press was not considered. A future study could look at the cumulative effect of CRI across four content areas using these rigorous strategies and ascertain this as the path for the effects of CareerStart.

Conclusion

The idea that teachers matter has some intuitive appeal because they are most proximal to the child. Because they primarily determine what goes on in their classrooms, teacher can ultimately decide on their own practices, including whether or not to implement mandated or recommended reforms (Kennedy, 2010; Wang et al., 2010). Teacher beliefs are culturally informed, largely by their own experiences as students (Pajares, 1992; Sykes, Bird & Kennedy, 2010) and reforms that do not accord with teacher belief systems often not only trigger passive resistance but active participation to change the new culture (Kelchtermans, 2005).

This may explain part of the appeal of changing teaching practice by changing the candidates in the teacher labor force to reflect these desired characteristics. More proximal strategies include changing how the teaching labor force is trained prior to service and supported during service. Worrisome trends in teacher recruitment and retention must be acknowledged when deciding which strategy to use. The teacher workforce has changed substantially over the last 30 years both in terms of average tenure and how teachers are prepared for teaching (Darling-Hammond et al. 2001; Glazerman et al, 2008; Gordon, Kane & Staiger, 2006). The trends in recruitment and retention, which may be associated with policies directed at teachers or students, are not positive (Clotfelder, Ladd, Vigdor & Diaz, 2004; Horng, 2009; Ladd, 2011), and higher salaries or bonuses may not be a panacea (Clotfelder et al, 2008a; 2008b). Trends in teacher workloads relative to other professions, including work on Sundays and holding second jobs, suggest that other professions may be more desirable (Krantz-Kent, 2008). Reforms that are integrated into teachers' existing practices may be more successful because they do not add to teachers' already burdened

schedules and demands (Valli & Buese, 2007). Without involvement of teachers, or attention to their beliefs, values, and needs, reform efforts may be stymied or have a demoralizing effect on both current teachers as well as reducing the pool of quality teachers by acting as disincentives on qualified candidates (Kelchtermans, 2005; Wang, Spalding, Odell, Klecka & Lin, 2010).

This study concerned the evidence for the theorized process for a program that lends support to teachers' efforts to promote engagement and learning in the classroom. The study demonstrated strong but inconclusive evidence that this process, career-relevant instruction, promoted student valuing. There is also weak evidence that engagement, as well as valuing, may have been affected by teachers' academic press. In sum, the evidence strongly suggests further investigations into the processes by which children learn. Because CareerStart has been shown to promote learning in an RCT, further investigations should be conducted to ascertain whether these effects are generalizable and if so, the processes by which they operate.

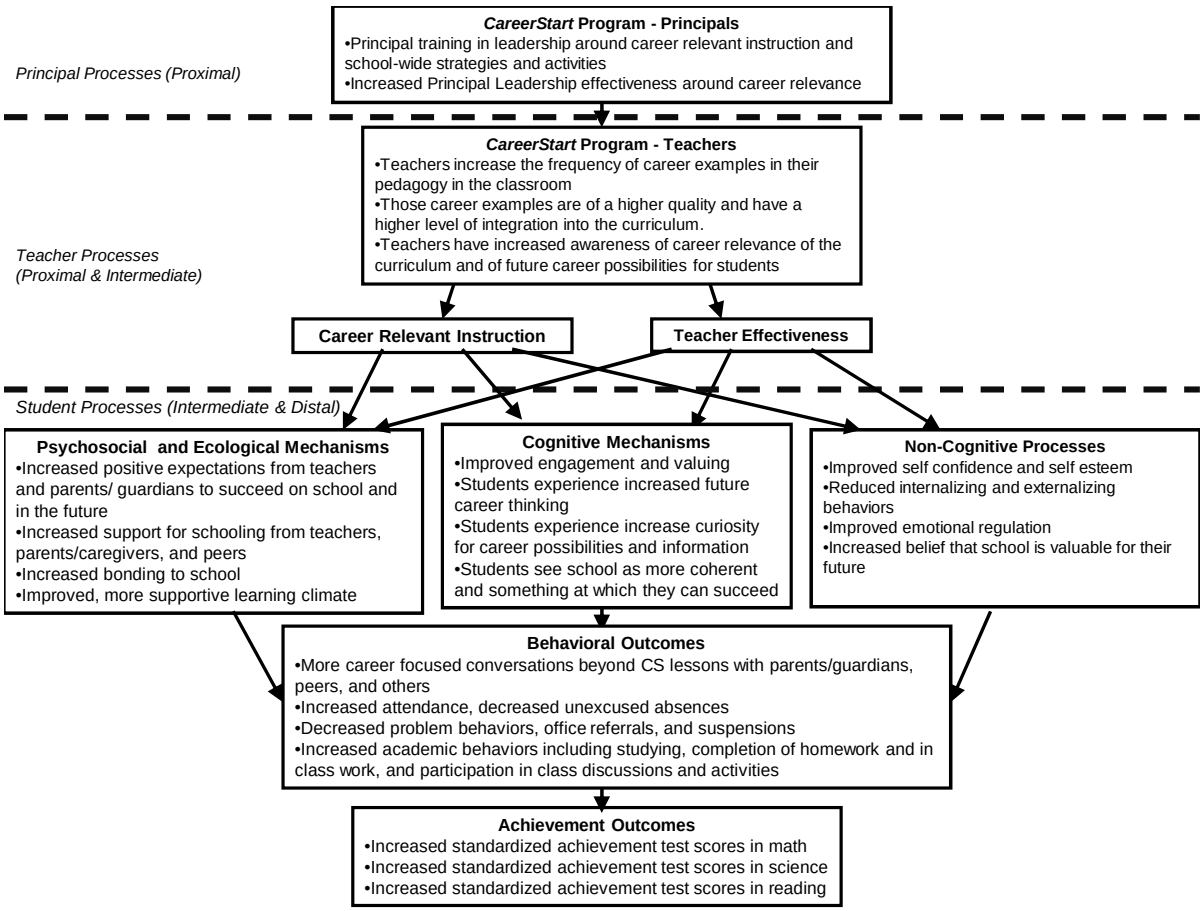

Figure 2.1. CareerStart theory of change model.

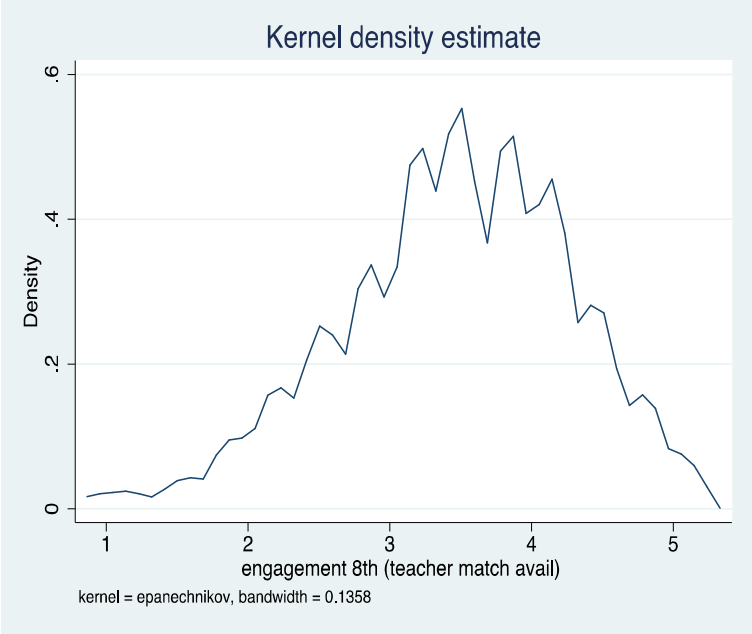

Figure 2.2. Kernel density engagement.

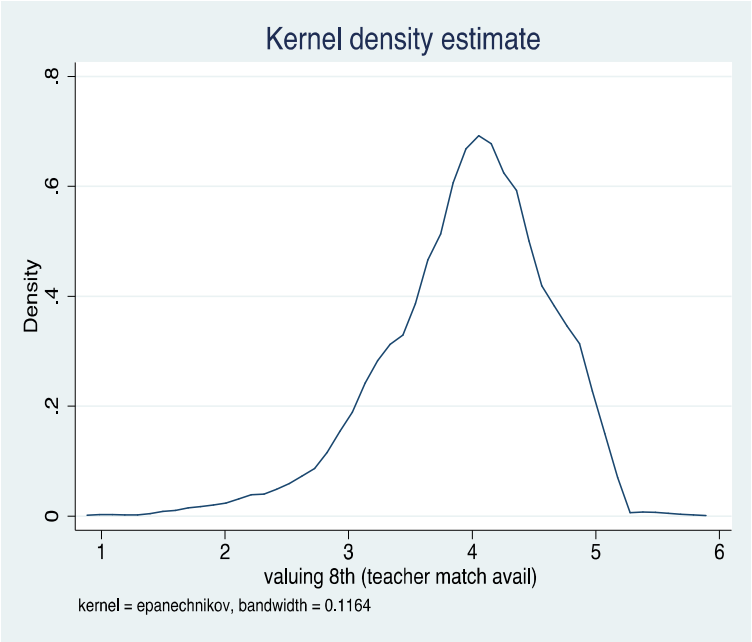

Figure 2.3. Kernel density valuing.

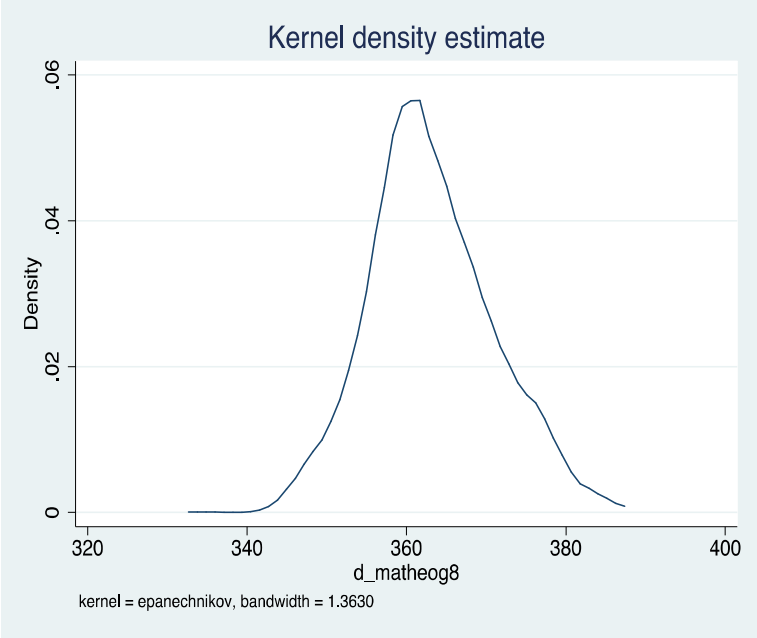

Figure 2.4. Kernel density math.

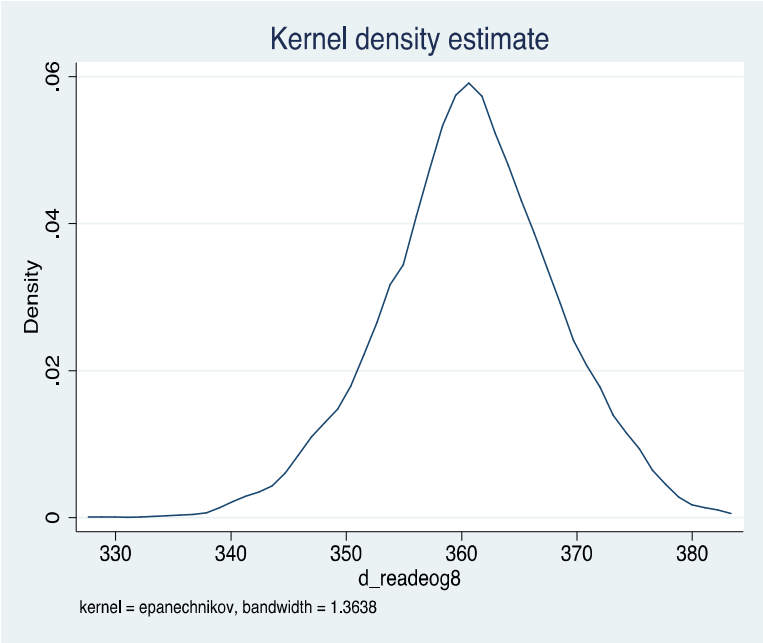

Figure 2.5. Kernel density reading.

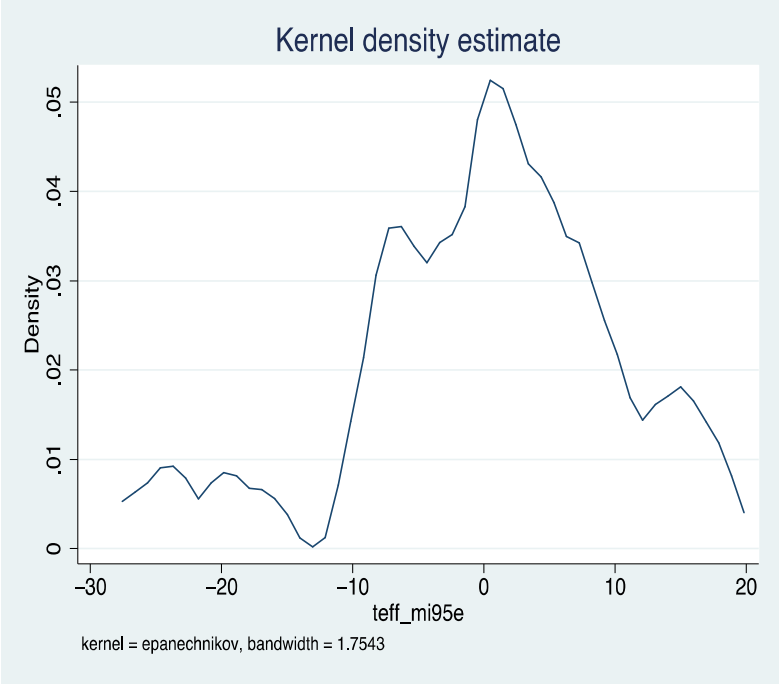

Figure 2.6. Kernel density LA CRI.

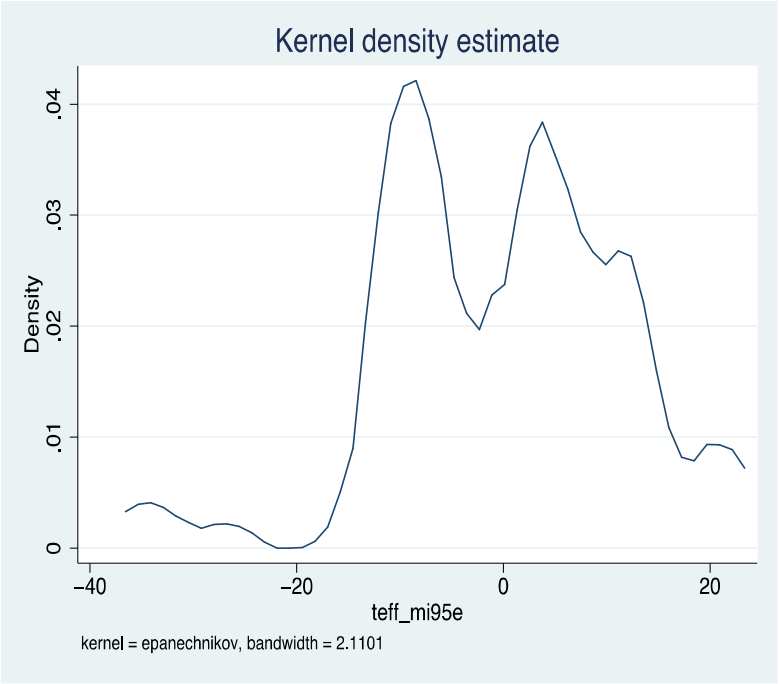

Figure 2.7. Kernel density math CRI.

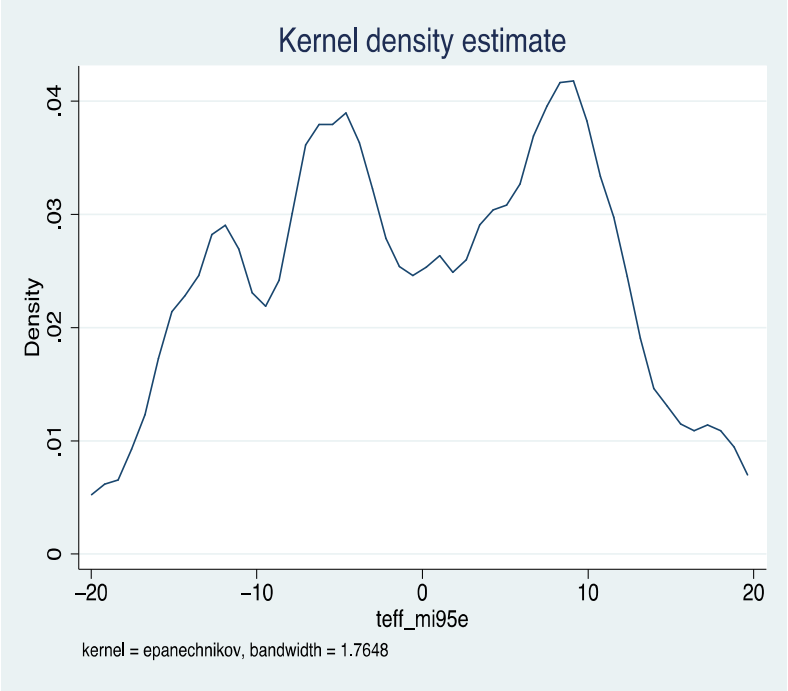

Figure 2.8. Kernel density science CRI.

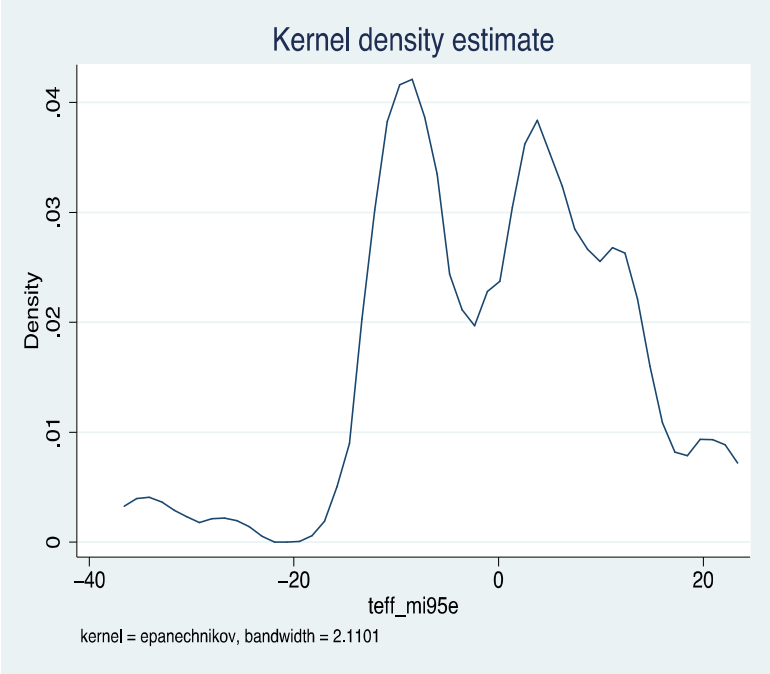

Figure 2.9. Kernel density social studies CRI.

Table 2.1.

Sample Characteristics (N = 2866)

Student Background, Home & Demographic Characteristics	**Mean/Prop**	**SD**	**Min**	**Median**	**Max**	**N(Missing)**
Black	0.30	0.46	0.00	0.00	1.00	0
Hispanic	0.15	0.36	0.00	0.00	1.00	0
Other R/E (Amer Ind/Asian/Aleutian/Pac Isle/Multi/Other)	0.06	0.24	0.00	0.00	1.00	0
Gender (female)	0.49	0.50	0.00	0.00	1.00	0
Receives free/reduced price lunches	0.52	0.50	0.00	1.00	1.00	0
Academically gifted	0.18	0.38	0.00	0.00	1.00	0
Enrolled in special education services	0.11	0.32	0.00	0.00	1.00	0
Parent marital status (single)	0.32	0.47	-0.37	0.00	1.26	0
Age	11.56	0.49	10.30	11.50	13.90	0
Student Test Score Performance Measures	**Mean**	**SD**	**Min**	**Median**	**Max**	**N(Missing)**
EOG score for math 8th grade	362.93	7.86	334.00	362.00	386.00	0
EOG score for math 7th grade	359.07	10.02	328.29	359.00	383.11	0
EOG score for math 6th grade	355.88	10.34	323.03	356.00	382.06	0
EOG score for math 5th grade	354.81	9.59	326.37	355.00	379.49	0
EOG score for math 4th grade	343.37	6.65	323.14	343.39	362.54	0
EOG score for math 3rd grade	339.31	4.95	323.28	339.44	353.94	0
EOG score for reading 8th grade	360.53	7.60	329.00	361.00	382.00	0
EOG score for reading 7th grade	355.43	9.62	325.36	356.00	379.00	0
EOG score for reading 6th grade	350.09	9.74	317.15	350.83	376.37	0
EOG score for reading 5th grade	347.63	8.92	317.11	347.51	370.87	0
EOG score for reading	342.42	9.89	309.55	343.07	368.19	0

4th grade EOG score for reading 3rd grade	338.11	10.02	303.78	338.63	365.84	0
Student Engagement and Psychosocial Measures	**Mean**	**SD**	**Min**	**Median**	**Max**	**N(Missing)**
Connectedness 8th grade	3.69	0.79	0.97	3.72	5.77	0
Connectedness 7th grade	3.79	0.80	1.00	4.00	5.94	0
Connectedness 6th grade	3.87	0.78	1.00	4.00	5.92	0
Engagement 8th grade	3.46	0.81	0.87	3.47	5.28	0
Engagement 7th grade	3.51	0.84	0.88	3.67	5.50	0
Engagement 6th grade	3.61	0.76	0.98	3.67	5.76	0
Valuing 8th grade	3.95	0.65	1.00	4.00	5.64	0
Valuing 7th grade	4.03	0.67	1.00	4.14	5.84	0
Valuing 6th grade	4.15	0.64	1.00	4.29	5.92	0
Career-Relevant Parenting/Home Career Relevance Measures by Response	**Mean**	**SD**	**Min**	**Median**	**Max**	**N(Missing)**
My parents/guardian show me the kind of things they do at work (SA or A = 1)	0.84	0.36	0.00	1.00	1.00	0
My parents/guardian tell me about their jobs (SA or A = 1)	0.89	0.32	0.00	1.00	1.00	0
My parents/guardian have shown me where they work (SA or A = 1)	0.95	0.23	0.00	1.00	1.00	0
My parents/guardian tell me about things that happen to them at work (SA or A = 1)	0.90	0.29	0.00	1.00	1.00	0
Future Preparation	**Mean**	**SD**	**Min**	**Median**	**Max**	**N(Missing)**
I plan to attend some type of college after high school (SA or A = 1)	0.97	0.18	0.00	1.00	1.00	0

Teacher Characteristics and Perspectives on CRI	Mean/Prop	SD	Min	Median	Max	N(Missing)
LA teacher reported perspective on CRI	34.31	6.37	17.00	35.00	44.00	127
LA teacher years of service	15.21	9.91	0.50	15.00	37.00	358
LA teacher gender (female)	0.84	0.37	0.00	1.00	1.00	127
LA teacher race (Black)	0.18	0.38	0.00	0.00	1.00	433
Math teacher reported perspective on CRI	36.22	6.16	22.00	37.00	44.00	413
Math teacher years of service	11.75	7.75	1.00	11.00	27.00	648
Math teacher gender (female)	0.76	0.43	0.00	1.00	1.00	483
Math teacher race (Black)	0.15	0.36	0.00	0.00	1.00	648
Science teacher reported perspective on CRI	35.75	6.44	19.00	38.00	44.00	170
Science teacher years of service	14.68	11.56	1.00	14.00	37.00	264
Science teacher gender (female)	0.63	0.48	0.00	1.00	1.00	170
Science teacher race (Black)	0.18	0.39	0.00	0.00	1.00	385
Social studies teacher reported perspective on CRI	34.80	5.11	24.00	34.00	44.00	117
Social studies teacher years of service	13.49	9.60	1.00	10.00	34.00	168
Social studies gender (female)	0.56	0.50	0.00	1.00	1.00	117
Social studies race (Black)	0.09	0.28	0.00	0.00	1.00	365

Teacher Quality and EVAAS Effectiveness Scores	Mean	SD	Min	Median	Max	N(Missing)
Teacher quality: academic press or rigor	11.45	2.25	2.99	12.00	17.86	0
Teacher quality: preparation for future schooling	11.86	2.37	3.00	12.00	18.50	0
Treatment Condition	**Mean**	**SD**	**Min**	**Median**	**Max**	**N(Missing)**
Enrolled in CareerStart treatment school in 8th grade	0.51	0.50	0.00	1.00	1.00	0

Table 2.2.

CRI Binary Variable

	Middle Category Set ("Uncertain") to Missing and Imputed					
	Mean	**SD**	**Min**	**Median**	**Max**	**N(Missing)**
Student-reported level of CRI, binary (SA or A = 1), 6th grade language arts	0.70	0.46	0.00	1.00	1.00	0
Student-reported level of CRI, binary (SA or A = 1), 7th grade language arts	0.66	0.47	0.00	1.00	1.00	0
Student-reported level of CRI, binary (SA or A = 1), 8th grade language arts	0.72	0.45	0.00	1.00	1.00	0
Student-reported level of CRI, binary (SA or A = 1), 6th grade math	0.68	0.47	0.00	1.00	1.00	0
Student-reported level of CRI, binary (SA or A = 1), 7th grade math	0.73	0.45	0.00	1.00	1.00	0
Student-reported level of CRI, binary (SA or A = 1), 8th grade math	0.70	0.46	0.00	1.00	1.00	0
Student-reported level of CRI, binary (SA or A = 1), 6th grade science	0.67	0.47	0.00	1.00	1.00	0
Student-reported level of CRI, binary (SA or A = 1), 7th grade science	0.70	0.46	0.00	1.00	1.00	0
Student-reported level of CRI, binary (SA or A = 1), 8th grade science	0.72	0.45	0.00	1.00	1.00	0
Student-reported level of CRI, binary (SA or A = 1), 6th grade social studies	0.65	0.48	0.00	1.00	1.00	0
Student-reported level of CRI, binary (SA or A = 1), 7th grade social studies	0.68	0.46	0.00	1.00	1.00	0
Student-reported level of CRI, binary (SA or A = 1), 8th grade social studies	0.69	0.46	0.00	1.00	1.00	0

Table 2.3.

Decomposition of Variance in CRI

Proportion of Variance at Teacher Level				
Using Logit				
	Teacher	Student	Total	Proportion at Teacher Level
Language Arts	0.30	3.29	3.59	0.08
Math	0.40	3.29	3.69	0.11
Science	0.29	3.29	3.58	0.08
Social Studies	0.20	3.29	3.49	0.06
Using LPM				
	Teacher	Student	Total	Proportion at Teacher Level
Language Arts	0.01	0.19	0.20	0.05
Math	0.02	0.20	0.21	0.07
Science	0.01	0.19	0.20	0.05
Social Studies	0.01	0.21	0.22	0.03

There are real differences between teachers (last column; from 3 to 11%, depending on the subject), but this means that 89 to 97% of the variance is within classroom; that is, between students.

Table 2.4.

CRI Measurement Model for Language Arts: Factors Explaining Variation in Student Perceptions of 8th Grade CRI

Language Arts	**OR**	**LO**	**SE**	**p value**	
Intercept (conditional mean for reference conditions)	3.18	1.16	0.126	0.000	***
Black	0.90	-0.11	0.149		
Hispanic	1.09	0.09	0.203	0.675	
Other R/E (Amer Ind/Asian/Aleutian/Pac Isle/Multi/Other)	0.81	-0.21	0.253	0.413	
Gender (female)	1.21	0.19	0.102	0.060	
Receives free/reduced price lunches	1.01	0.01	0.157	0.935	
Academically gifted	0.96	-0.04	0.195	0.834	
Enrolled in special education services	0.94	-0.07	0.225	0.771	
Age	1.09	0.09	0.170	0.598	
Age-Squared	0.86	-0.15	0.229	0.500	
Age-Cubed	1.16	0.15	0.183	0.430	
Parent marital status (single)	1.03	0.03	0.134	0.814	
EOG score for reading 5th grade	1.00	0.00	0.016	0.786	
EOG score for reading 5th grade-Squared	1.00	0.00	0.001	0.970	
EOG score for reading 5th grade-Cubed	1.00	0.00	0.000	0.582	
EOG score for reading 4th grade	0.98	-0.02	0.018	0.404	
EOG score for reading 4th grade-Squared	1.00	0.00	0.001	0.476	
EOG score for reading 4th grade-Cubed	1.00	0.00	0.000	0.477	
EOG score for reading 3rd grade	0.99	-0.01	0.016	0.684	
EOG score for reading 3rd grade-Squared	1.00	0.00	0.001	0.477	
EOG score for reading 3rd grade-Cubed	1.00	0.00	0.000	0.326	
Student-reported level of CRI, binary (SA or A = 1), 6th grade language arts	1.57	0.45	0.116	0.000	***
Student-reported level of CRI, binary (SA or A = 1), 7th grade language arts	2.56	0.94	0.129	0.000	***
My parents/guaridan tell me about the kind of work they do (SA or A = 1)	1.15	0.14	0.274	0.621	
My parents/guardian show me the kind of things they do at work (SA or	1.58	0.46	0.171	0.008	**

A = 1)					
My parents/guardian tell me about their jobs (SA or A = 1)	1.19	0.17	0.309	0.582	
My parents/guardian have shown me where they work (SA or A = 1)	1.26	0.23	0.274	0.406	
My parents/guardian tell me about things that happen to them at work (SA or A = 1)	1.30	0.26	0.201	0.192	
I plan to attend some type of college after high school (SA or A = 1)	2.79	1.03	0.285	0.000	***

*** p < .001; ** p < .01; * p < .05

Table 2.5.

CRI Measurement Model for Math: Factors Explaining Variation in Student Perceptions of 8th Grade CRI

Math	**OR**	**LO**	**SE**	**p value**	
Intercept (conditional mean for reference cond.)	2.70	0.99	0.146	0.000	***
Black	1.12	0.11	0.148	0.456	
Hispanic	1.49	0.40	0.188	0.034	*
Other R/E (Amer Ind/Asian/Aleutian/Pac Isle/Multi/Other)	0.93	-0.07	0.225	0.743	
Gender (female)	0.94	-0.06	0.120	0.636	
Receives free/reduced price lunches	1.17	0.15	0.144	0.289	
Academically gifted	1.30	0.26	0.185	0.158	
Enrolled in special education services	0.90	-0.11	0.190	0.569	
Age	1.14	0.13	0.159	0.422	
Age-Squared	0.96	-0.04	0.245	0.855	
Age-Cubed	1.02	0.02	0.193	0.901	
Parent marital status (single)	0.96	-0.04	0.122	0.733	
EOG score for math 5th grade	1.01	0.01	0.017	0.731	
EOG score for math 5th grade-Squared	1.00	0.00	0.001	0.852	
EOG score for math 5th grade-Cubed	1.00	0.00	0.000	0.722	
EOG score for math 4th grade	0.98	-0.02	0.023	0.417	
EOG score for math 4th grade-Squared	1.00	0.00	0.002	0.375	
EOG score for math 4th grade-Cubed	1.00	0.00	0.000	0.770	
EOG score for math 3rd grade	1.02	0.02	0.034	0.476	
EOG score for math 3rd grade-Squared	1.00	0.00	0.003	0.451	
EOG score for math 3rd grade-Cubed	1.00	0.00	0.000	0.630	
Student-reported level of CRI, binary (SA or A = 1), 6th grade math	1.84	0.61	0.166	0.001	**
Student-reported level of CRI, binary (SA or A = 1), 7th grade math	2.80	1.03	0.131	0.000	***
My parents/guaridan tell me about the kind of work they do (SA or A = 1)	1.90	0.64	0.296	0.037	*
My parents/guardian show me the kind of things they do at work (SA or A = 1)	1.44	0.36	0.177	0.042	*
My parents/guardian tell me about their jobs (SA or A = 1)	0.67	-0.40	0.301	0.195	
My parents/guardian have shown me where they work (SA or A = 1)	1.45	0.37	0.261	0.155	
My parents/guardian tell me about things that happen to them at work (SA or A = 1)	1.47	0.38	0.210	0.069	
I plan to attend some type of college after high school (SA or A = 1)	2.58	0.95	0.266	0.000	***

*** p < .001; ** p < .01; * p < .05

Table 2.6.

CRI Measurement Model for Science: Factors Explaining Variation in Student Perceptions of 8th Grade CRI

Science	**OR**	**LO**	**SE**	**p value**	
Intercept (conditional mean for reference conditions)	3.15	1.15	0.129	0.000	***
Black	1.14	0.13	0.151	0.404	
Hispanic	1.45	0.37	0.194	0.057	
Other R/E (Amer Ind/Asian/Aleutian/Pac Isle/Multi/Other)	1.52	0.42	0.260	0.112	
Gender (female)	0.94	-0.07	0.114	0.564	
Receives free/reduced price lunches	1.13	0.13	0.153	0.414	
Academically gifted	1.26	0.23	0.188	0.215	
Enrolled in special education services	1.53	0.43	0.237	0.078	
Age	1.20	0.19	0.167	0.269	
Age-Squared	0.81	-0.21	0.221	0.340	
Age-Cubed	0.98	-0.02	0.171	0.898	
Parent marital status (single)	1.02	0.02	0.131	0.890	
EOG score for math 5th grade	1.02	0.02	0.013	0.081	
EOG score for math 5th grade-Squared	1.00	0.00	0.001	0.763	
EOG score for math 4th grade	0.98	-0.02	0.018	0.201	
EOG score for math 4th grade-Squared	1.00	0.00	0.002	0.705	
EOG score for math 3rd grade	0.98	-0.02	0.028	0.432	
EOG score for math 3rd grade-Squared	1.00	0.00	0.003	0.684	
EOG score for reading 5th grade	1.00	0.00	0.014	0.853	
EOG score for reading 5th grade-Squared	1.00	0.00	0.001	0.511	
EOG score for reading 4th grade	1.00	0.00	0.011	0.795	
EOG score for reading 4th grade-Squared	1.00	0.00	0.001	0.396	
EOG score for reading 3rd grade	1.01	0.01	0.014	0.559	
EOG score for reading 3rd grade-Squared	1.00	0.00	0.001	0.260	
Student-reported level of CRI, binary (SA or A = 1), 6th grade science	1.47	0.39	0.126	0.003	**
Student-reported level of CRI, binary (SA or A = 1), 7th grade science	1.87	0.63	0.115	0.000	***
My parents/guaridan tell me about the kind of work they do (SA or A = 1)	1.08	0.08	0.259	0.763	
My parents/guardian show me the kind of things they do at work (SA or A = 1)	1.68	0.52	0.221	0.026	*
My parents/guardian tell me about their jobs (SA or A = 1)	1.03	0.03	0.264	0.898	

My parents/guardian have shown me where they work (SA or A = 1)	1.29	0.26	0.290	0.382	
My parents/guardian tell me about things that happen to them at work (SA or A = 1)	1.68	0.52	0.210	0.015	*
I plan to attend some type of college after high school (SA or A = 1)	2.32	0.84	0.262	0.001	**

*** p < .001; ** p < .01; * p < .05

Table 2.7.

CRI Measurement Model for Social Studies: Factors Explaining Variation in Student Perceptions of 8th Grade CRI

Social Studies	**OR**	**LO**	**SE**	**p value**	
Intercept (conditional mean for reference conditions)	2.34	0.85	0.122	0.000	***
Black	1.01	0.01	0.153	0.939	
Hispanic	1.15	0.14	0.189	0.474	
Other R/E (Amer Ind/Asian/Aleutian/Pac Isle/Multi/Other)	1.08	0.08	0.218	0.710	
Gender (female)	0.94	-0.07	0.111	0.550	
Receives free/reduced price lunches	1.37	0.32	0.134	0.019	*
Academically gifted	1.15	0.14	0.186	0.465	
Enrolled in special education services	1.19	0.18	0.209	0.401	
Age	1.07	0.07	0.176	0.689	
Age-Squared	1.02	0.02	0.241	0.950	
Age-Cubed	1.05	0.05	0.191	0.799	
Parent marital status (single)	0.98	-0.02	0.116	0.897	
EOG score for math 5th grade	1.01	0.01	0.016	0.657	
EOG score for math 5th grade-Squared	1.00	0.00	0.001	0.618	
EOG score for math 4th grade	1.00	0.00	0.017	0.952	
EOG score for math 4th grade-Squared	1.00	0.00	0.002	0.569	
EOG score for math 3rd grade	1.01	0.01	0.026	0.638	
EOG score for math 3rd grade-Squared	1.00	0.00	0.003	0.952	
EOG score for reading 5th grade	1.01	0.01	0.013	0.248	
EOG score for reading 5th grade-Squared	1.00	0.00	0.001	0.342	
EOG score for reading 4th grade	0.98	-0.02	0.011	0.035	*
EOG score for reading 4th grade-Squared	1.00	0.00	0.001	0.162	
EOG score for reading 3rd grade	0.99	-0.01	0.012	0.548	
EOG score for reading 3rd grade-Squared	1.00	0.00	0.001	0.680	
Student-reported level of CRI, binary (SA or A = 1), 6th grade social studies	1.45	0.37	0.116	0.002	**
Student-reported level of CRI, binary (SA or A = 1), 7th grade social studies	2.25	0.81	0.104	0.000	***
My parents/guaridan tell me about the kind of work they do (SA or A = 1)	1.31	0.27	0.242	0.270	
My parents/guardian show me the kind of things they do at work (SA or A = 1)	1.50	0.40	0.186	0.034	*
My parents/guardian tell me about their jobs (SA or A = 1)	1.10	0.09	0.250	0.717	

My parents/guardian have shown me where they work (SA or A = 1)	0.79	-0.23	0.287	0.422	
My parents/guardian tell me about things that happen to them at work (SA or A = 1)	1.45	0.37	0.218	0.094	
I plan to attend some type of college after high school (SA or A = 1)	2.49	0.91	0.295	0.003	**

*** p < .001; ** p < .01; * p < .05

Table 2.8.

IVE Models without Teacher-Level Imputation, Different Covariate Specifications

		Uncertain as Missing (and Imputed)	
Dependent Variable	**CRI Subject**	**Indv, Teacher Assignment, School[1]**	**Indv, All Teacher, School[2]**
8th Grade Math	Math		
8th Grade Reading	LA		
Engagement	Math		
	LA		
	Science		
	Social Studies		
Valuing	Math		
	LA	.01**	
	Science	.01***	
	Social Studies	.01**	.01*

*** p < .001; ** p < .01; * p < .05

[1]Number of teachers: math, 48; language arts, 45; science, 50; social studies, 42.

[2]Number of teachers: math, 33; language arts, 34; science, 33; social studies, 36.

Table 2.9.

IVE Models with Teacher-Level Imputation, Different Covariate Specifications

Dependent Variable	**CRI Subject**	**Indv, Teacher Assignment, School[1]**	**Indv, All Teacher, School[2]**
8th Grade Math	Math		
8th Grade Reading	LA		
Engagement	Math		
	LA		
	Science		
	Social Studies		
Valuing	Math	.01*	.01*
	LA	.01**	.01***
	Science	.01***	.01***
	Social Studies	.01**	.01**

*** p < .001; ** p < .01; * p < .05

[1]Number of teachers: math, 48; language arts, 45; science, 50; social studies, 42.

[2]Number of teachers: math, 33; language arts, 34; science, 33; social studies, 36.

Table 2.10.

IVE Main Models with Observations Having Missing Teacher Covariates Deleted

Dependent Variable	**CRI Subject**	**Indv, Teacher Assignment, School[1]**	**Indv, All Teacher, School[2]**
8th Grade Math	Math		
8th Grade Reading	LA		
Engagement	Math		
	LA		
	Science		
	Social Studies		
Valuing	Math	.01*	
	LA	.01*	
	Science		
	Social Studies		.01*

*** p < .001; ** p < .01; * p < .05

[1]Number of teachers: math, 48; language arts, 45; science, 50; social studies, 42.

[2]Number of teachers: math, 33; language arts, 34; science, 33; social studies, 36.

Table 2.11.

No IVE: Three-Level Rich Covariate Models without Teacher Level Imputation, Different Covariate Specifications

Dependent Variable	**CRI Subject**	**Indv, Teacher Assignment, School[1]**	**Indv, All Teacher, School[2]**
8th Grade Math	Math	.04*	
8th Grade Reading	LA		
Engagement	Math		
	LA		
	Science		
	Social Studies		
Valuing	Math	.01*	.01*
	LA	.01*	
	Science	.01*	.01*
	Social Studies	.01**	

*** p < .001; ** p < .01; * p < .05

[1]Number of teachers: math, 48; language arts, 45; science, 50; social studies, 42.

[2]Number of teachers: math, 33; language arts, 34; science, 33; social studies, 36.

Table 2.12.

IVE Models without Teacher Level Imputation, Different Covariate Specifications

		Uncertain as Disagree	
Dependent Variable	**CRI Subject**	**Indv, Teacher Assignment, School[1]**	**Indv, All Teacher, School[2]**
8th Grade Math	Math		
8th Grade Reading	LA		
Engagement	Math		
	LA		
	Science		
	Social Studies		
Valuing	Math		.01*
	LA	.004*	
	Science	.004**	.005*
	Social Studies	.01**	.005*

*** p < .001; ** p < .01; * p < .05

[1]Number of teachers: math, 48; language arts, 45; science, 50; social studies, 42.

[2]Number of teachers: math, 33; language arts, 34; science, 33; social studies, 36.

Table 2.13.

EOG Math 8th Grade IVE/LATE

	Main: All Individual Covariates, Teacher Assignment, School Level Controls[1]				Teacher Covs: All Individual Covariates, Teacher Assignment, School Level Controls; with Teacher Covariates[2]			
Variable	**Est.**	**SE**	**t value**		**Est.**	**SE**	**t value**	
Intercept (conditional mean for reference conditions)	113.34	35.27	3.213	**	130.66	963.45	0.136	
Independent Variable/Treatment Effect								
CRI (Student-reported probability that teacher frequenty used CRI)	0.03	0.02	1.679		0.00	0.67	-0.006	
Pretests (ES and MS)								
Math EOG 7th grade	0.32	0.02	14.036	***	0.35	0.03	12.439	***
Math EOG 7th grade-Squared	0.00	0.00	3.451	***	0.00	0.00	3.501	***
Math EOG 7th grade-Cubed	0.00	0.00	-0.691		0.00	0.00	-0.752	
Math EOG 6th grade	0.21	0.02	9.245	***	0.21	0.03	7.748	***
Math EOG 6th grade-Squared	0.00	0.00	0.869		0.00	0.00	0.790	
Math EOG 6th grade-Cubed	0.00	0.00	1.940		0.00	0.00	1.688	
Math EOG 5th grade	0.05	0.03	1.851		0.05	0.03	1.496	
Math EOG 5th grade-Squared	0.00	0.00	0.950		0.00	0.00	1.315	
Math EOG 5th grade-Cubed	0.00	0.00	0.217		0.00	0.00	0.148	
Math EOG 4th grade	0.02	0.03	0.598		0.03	0.04	0.692	
Math EOG 4th grade-Squared	0.00	0.00	-0.955		0.00	0.00	-0.784	
Math EOG 4th grade-Cubed	0.00	0.00	2.120	*	0.00	0.00	1.370	
Math EOG 3rd grade	0.06	0.04	1.324		0.04	0.05	0.733	
Math EOG 3rd grade-Squared	0.00	0.00	-0.325		0.00	0.00	-0.684	

Math EOG 3rd grade-Cubed	0.00	0.00	-0.315		0.00	0.00	0.005	
Demographics and Family Variables								
Age	-0.61	0.20	-3.029	**	-0.52	0.23	-2.207	*
Age-Squared	-0.56	0.31	-1.785		-0.52	0.36	-1.423	
Age-Cubed	0.43	0.22	1.961	*	0.39	0.25	1.551	
Black	0.45	0.20	2.279	*	0.52	0.23	2.274	*
Hispanic	0.58	0.24	2.382	*	0.49	0.28	1.767	
Other R/E (Amer Ind/Asian/Aleutian/Pac Isle/Multi/Other)	1.10	0.29	3.732	***	1.29	0.36	3.586	***
Gender (female)	0.38	0.14	2.787	**	0.43	0.16	2.701	**
Receives free/reduced price lunches	-0.36	0.18	-1.954		-0.37	0.21	-1.721	
Academically gifted	0.76	0.24	3.128	**	0.73	0.29	2.486	*
Enrolled in special education services	-0.43	0.25	-1.697		-0.68	0.29	-2.390	*
Parent marital status (single)	0.04	0.16	0.260		0.00	0.19	0.015	
Teacher and School Characteristics								
Math teacher years of service					-0.02	0.99	-0.018	
Math teacher gender: female					0.57	19.08	0.030	
Math teacher race: Black					-0.26	16.66	-0.016	
Avg 7th grade math score of students with this 8th grade math teacher	0.74	0.05	14.035	***	0.70	1.34	0.523	
School average free/reduced lunch	-0.85	1.67	-0.507		-1.57	43.11	-0.036	
School average math 7th grade	-0.05	0.07	-0.650		-0.05	2.10	-0.024	

*** $p < .001$; ** $p < .01$; * $p < .05$

[1]Number of teachers: math, 48; language arts, 45; science, 50; social studies, 42.

[2]Number of teachers: math, 33; language arts, 34; science, 33; social studies, 36.

Table 2.14.

EOG Reading 8th Grade IVE/LATE

	Main: All Individual Covariates, Teacher Assignment, School Level Controls[1]				Teacher Covs: All Individual Covariates, Teacher Assignment, School Level Controls; with Teacher Covariates[2]			
Variable	**Est.**	**SE**	**t value**		**Est.**	**SE**	**t value**	
Intercept (conditional mean for reference conditions)	77.34	21.94	3.525	***	112.85	1308.27	0.086	
Independent Variable/Treatment Effect								
CRI (Student-reported probability that teacher frequenty used CRI)	0.02	0.01	1.106		0.01	0.70	0.013	
Pretests (ES and MS)								
Reading EOG 7th grade	0.24	0.02	10.873	***	0.22	0.03	8.690	***
Reading EOG 7th grade-Squared	0.00	0.00	0.495		0.00	0.00	0.709	
Reading EOG 7th grade-Cubed	0.00	0.00	1.093		0.00	0.00	2.853	**
Reading EOG 6th grade	0.14	0.03	5.432	***	0.15	0.03	4.944	***
Reading EOG 6th grade-Squared	0.00	0.00	-1.237		0.00	0.00	-1.367	
Reading EOG 6th grade-Cubed	0.00	0.00	-0.106		0.00	0.00	-0.710	
Reading EOG 5th grade	0.14	0.03	5.206	***	0.12	0.03	4.308	***
Reading EOG 5th grade-Squared	0.00	0.00	1.481		0.00	0.00	1.697	
Reading EOG 5th grade-Cubed	0.00	0.00	0.014		0.00	0.00	0.699	
Reading EOG 4th grade	0.13	0.02	5.577	***	0.14	0.02	5.739	***
Reading EOG 4th grade-Squared	0.00	0.00	-0.039		0.00	0.00	-0.606	

Reading EOG 4th grade-Cubed	0.00	0.00	-1.183		0.00	0.00	-1.734	
Reading EOG 3rd grade	0.06	0.02	2.684	**	0.05	0.02	2.010	*
Reading EOG 3rd grade-Squared	0.00	0.00	0.135		0.00	0.00	0.594	
Reading EOG 3rd grade-Cubed	0.00	0.00	0.506		0.00	0.00	0.609	
Demographics and Family Variables								
Age	-0.42	0.21	-1.985	*	-0.27	0.25	-1.095	
Age-Squared	-0.49	0.33	-1.474		-0.63	0.37	-1.689	
Age-Cubed	0.06	0.24	0.241		0.00	0.29	-0.012	
Black	-0.38	0.21	-1.827		-0.31	0.23	-1.338	
Hispanic	-0.19	0.26	-0.729		-0.01	0.29	-0.031	
Other R/E (Amer Ind/Asian/Aleutian/Pac Isle/Multi/Other)	0.03	0.31	0.105		0.31	0.36	0.852	
Gender (female)	-0.21	0.14	-1.459		-0.26	0.16	-1.656	
Receives free/reduced price lunches	-0.35	0.19	-1.783		-0.35	0.21	-1.669	
Academically gifted	0.27	0.26	1.073		0.19	0.28	0.665	
Enrolled in special education services	-0.05	0.29	-0.186		-0.29	0.33	-0.880	
Parent marital status (single)	0.04	0.17	0.209		-0.06	0.19	-0.322	
Teacher and School Characteristics								
LA teacher years of service					0.00	0.61	-0.004	
LA teacher gender: female					-0.35	16.30	-0.021	
LA teacher race: Black					-0.09	15.44	-0.006	
Avg 7th grade reading score of students with this 8th grade LA teacher	0.81	0.04	19.384	***	0.75	2.04	0.368	
School average free/reduced lunch	0.36	0.97	0.375		-0.67	48.79	-0.014	
School average reading 7th grade	-0.02	0.05	-0.382		-0.05	2.30	-0.022	

*** p < .001; ** p < .01; * p < .05;

[1]Number of teachers: math, 48; language arts, 45; science, 50; social studies, 42;

[2]Number of teachers: math, 33; language arts, 34; science, 33; social studies, 36.

Table 2.15.

Engagement, Math CRI IVE/LATE

	Main: All Individual Covariates, Teacher Assignment, School Level Controls[1]				Teacher Covs: All Individual Covariates, Teacher Assignment, School Level Controls; with Teacher Covariates[2]			
Variable	**Est.**	**SE**	**t value**		**Est.**	**SE**	**t value**	
Intercept (conditional mean for reference conditions)	-5.45	3.21	-1.701		-4.55	4.12	-1.105	
Independent Variable/Treatment Effect								
CRI (Student-reported probability that teacher frequenty used CRI)	0.00	0.00	-0.073		0.00	0.00	0.375	
Pretests (ES and MS)								
Engagement, 7th grade	0.50	0.02	21.921	***	0.51	0.03	19.863	***
Engagement 7th grade-Squared	0.06	0.01	4.296	***	0.05	0.02	3.406	***
Engagement, 6th grade	0.18	0.02	7.137	***	0.18	0.03	6.414	***
Engagement 6th grade-Squared	-0.02	0.02	-1.294		-0.02	0.02	-0.751	
Math EOG 5th grade	0.00	0.00	-0.319		0.00	0.00	0.015	
Reading EOG 5th grade	0.00	0.00	0.267		0.00	0.00	0.505	
Math EOG 4th grade	0.00	0.00	0.100		0.00	0.01	0.144	
Reading EOG 4th grade	0.00	0.00	0.757		0.00	0.00	0.563	
Math EOG 3rd grade	0.00	0.01	0.734		0.01	0.01	1.119	
Reading EOG 3rd grade	0.00	0.00	-0.915		0.00	0.00	-1.140	
Demographics and Family Variables								
Age	0.03	0.04	0.780		0.05	0.04	1.009	

Age-Squared	0.06	0.06	1.140		0.06	0.06	0.977	
Age-Cubed	-0.06	0.04	-1.376		-0.05	0.04	-1.085	
Black	0.12	0.04	3.507	***	0.15	0.04	3.621	***
Hispanic	0.05	0.05	1.194		0.07	0.05	1.294	
Other R/E (Amer Ind/Asian/Aleutian/Pac Isle/Multi/Other)	0.01	0.06	0.164		0.08	0.07	1.156	
Gender (female)	0.04	0.03	1.420		0.02	0.03	0.532	
Receives free/reduced price lunches	-0.13	0.04	-3.641	***	-0.10	0.04	-2.676	**
Academically gifted	0.03	0.04	0.658		-0.01	0.05	-0.151	
Enrolled in special education services	-0.08	0.05	-1.545		-0.09	0.05	-1.695	
Parent marital status (single)	-0.06	0.03	-1.990	*	-0.04	0.03	-1.220	
Teacher and School Characteristics								
Math teacher years of service					0.00	0.01	-0.091	
Math teacher gender: female					0.03	0.11	0.295	
Math teacher race: Black					0.02	0.11	0.150	
Avg 7th grade math score of students with this 8th grade math teacher	0.02	0.01	2.811	**	0.02	0.01	1.957	
School average free/reduced lunch	0.28	0.23	1.228		0.10	0.25	0.393	

*** p < .001; ** p < .01; * p < .05;
[1]Number of teachers: math, 48; language arts, 45; science, 50; social studies, 42;
[2]Number of teachers: math, 33; language arts, 34; science, 33; social studies, 36.

Table 2.16.

Engagement, LA CRI IVE/LATE

	Main: All Individual Covariates, Teacher Assignment, School Level Controls[1]				Teacher Covs: All Individual Covariates, Teacher Assignment, School Level Controls; with Teacher Covariates[2]			
Variable	**Est.**	**SE**	**t value**		**Est.**	**SE**	**t value**	
Intercept (conditional mean for reference conditions)	-6.46	2.82	-2.286	*	-11.91	3.90	-3.052	**
Independent Variable/Treatment Effect								
CRI (Student-reported probability that teacher frequenty used CRI)	0.00	0.00	1.177		0.00	0.00	0.217	
Pretests (ES and MS)								
Engagement, 7th grade	0.50	0.02	21.857	***	0.50	0.02	20.761	***
Engagement 7th grade-Squared	0.05	0.01	3.959	***	0.06	0.02	3.923	***
Engagement, 6th grade	0.18	0.02	7.318	***	0.19	0.02	7.476	***
Engagement 6th grade-Squared	-0.02	0.02	-1.220		-0.03	0.02	-1.271	
Math EOG 5th grade	0.00	0.00	-0.257		0.00	0.00	-0.204	
Reading EOG 5th grade	0.00	0.00	0.082		0.00	0.00	0.367	
Math EOG 4th grade	0.00	0.00	0.091		0.00	0.01	0.040	
Reading EOG 4th grade	0.00	0.00	0.614		0.00	0.00	0.365	
Math EOG 3rd grade	0.01	0.01	0.777		0.01	0.01	0.837	
Reading EOG 3rd grade	0.00	0.00	-0.948		0.00	0.00	-1.015	
Demographics and Family Variables								
Age	0.02	0.04	0.581		0.04	0.05	0.761	

Age-Squared	0.07	0.06	1.180		0.08	0.06	1.231	
Age-Cubed	-0.06	0.04	-1.351		-0.06	0.05	-1.263	
Black	0.12	0.04	3.240	**	0.14	0.04	3.385	***
Hispanic	0.05	0.05	1.084		0.04	0.05	0.844	
Other R/E (Amer Ind/Asian/Aleutian/Pac Isle/Multi/Other)	0.00	0.06	-0.016		0.02	0.06	0.265	
Gender (female)	0.04	0.03	1.495		0.02	0.03	0.809	
Receives free/reduced price lunches	-0.13	0.04	-3.623	***	-0.12	0.04	-3.185	**
Academically gifted	0.06	0.05	1.336		0.06	0.05	1.162	
Enrolled in special education services	-0.09	0.05	-1.701		-0.07	0.06	-1.158	
Parent marital status (single)	-0.05	0.03	-1.812		-0.05	0.03	-1.657	
Teacher and School Characteristics								
LA teacher years of service					0.00	0.00	0.923	
LA teacher gender: female					0.00	0.07	-0.056	
LA teacher race: Black					0.17	0.07	2.446	*
Avg 7th grade reading score of students with this 8th grade LA teacher	0.03	0.01	3.542	***	0.04	0.01	3.940	***
School average free/reduced lunch	0.25	0.16	1.546		0.40	0.18	2.265	*

*** p < .001; ** p < .01; * p < .05;
[1]Number of teachers: math, 48; language arts, 45; science, 50; social studies, 42;
[2]Number of teachers: math, 33; language arts, 34; science, 33; social studies, 36.

Table 2.17.

Engagement, Science CRI IVE/LATE

	Main: All Individual Covariates, Teacher Assignment, School Level Controls[1]				Teacher Covs: All Individual Covariates, Teacher Assignment, School Level Controls; with Teacher Covariates[2]			
Variable	**Est.**	**SE**	**t value**		**Est.**	**SE**	**t value**	
Intercept (conditional mean for reference conditions)	1.22	1.60	0.766		2.51	1.56	1.608	
Independent Variable/Treatment Effect								
CRI (Student-reported probability that teacher frequenty used CRI)	0.00	0.00	1.257		0.005	0.00	1.447	
Pretests (ES and MS)								
Engagement, 7th grade	0.50	0.02	22.508	***	0.50	0.02	19.820	***
Engagement 7th grade-Squared	0.05	0.01	4.043	***	0.06	0.01	3.871	***
Engagement, 6th grade	0.17	0.02	7.064	***	0.18	0.03	6.812	***
Engagement 6th grade-Squared	-0.03	0.02	-1.416		-0.03	0.02	-1.407	
Math EOG 5th grade	0.00	0.00	-0.240		0.00	0.00	-0.010	
Reading EOG 5th grade	0.00	0.00	0.181		0.00	0.00	0.081	
Math EOG 4th grade	0.00	0.00	0.167		0.00	0.00	-0.214	
Reading EOG 4th grade	0.00	0.00	0.671		0.00	0.00	1.140	
Math EOG 3rd grade	0.01	0.01	0.843		0.00	0.01	0.499	
Reading EOG 3rd grade	0.00	0.00	-0.918		0.00	0.00	-0.971	
Demographics and Family Variables								
Age	0.03	0.04	0.615		0.01	0.05	0.267	

Age-Squared	0.07	0.06	1.240		0.06	0.06	1.059	
Age-Cubed	-0.05	0.04	-1.330		-0.04	0.04	-0.903	
Black	0.13	0.04	3.575	***	0.13	0.04	3.326	***
Hispanic	0.05	0.05	1.083		0.05	0.05	0.910	
Other R/E (Amer Ind/Asian/Aleutian/Pac Isle/Multi/Other)	0.01	0.06	0.099		0.00	0.06	-0.051	
Gender (female)	0.04	0.03	1.328		0.03	0.03	0.924	
Receives free/reduced price lunches	-0.13	0.04	-3.544	***	-0.14	0.04	-3.602	***
Academically gifted	0.01	0.04	0.202		0.00	0.05	-0.008	
Enrolled in special education services	-0.09	0.05	-1.893		-0.10	0.05	-1.968	*
Parent marital status (single)	-0.06	0.03	-1.976	*	-0.06	0.03	-1.918	
Teacher and School Characteristics								
Science teacher years of service					0.00	0.00	-0.509	
Science teacher gender: female					0.01	0.06	0.161	
Science teacher race: Black					0.04	0.08	0.469	
Avg 7th grade math score of students with same 8th grade math teacher	0.01	0.00	1.431		0.00	0.00	0.688	
School average free/reduced lunch	-0.06	0.11	-0.526		-0.17	0.14	-1.258	

*** p < .001; ** p < .01; * p < .05;
[1]Number of teachers: math, 48; language arts, 45; science, 50; social studies, 42;
[2]Number of teachers: math, 33; language arts, 34; science, 33; social studies, 36.

Table 2.18.

Engagement, SS CRI IVE/LATE

	Main: All Individual Covariates, Teacher Assignment, School Level Controls[1]				Teacher Covs: All Individual Covariates, Teacher Assignment, School Level Controls; with Teacher Covariates[2]			
Variable	**Est.**	**SE**	**t value**		**Est.**	**SE**	**t value**	
Intercept (conditional mean for reference conditions)	1.71	1.52	1.120		1.79	1.69	1.059	
Independent Variable/Treatment Effect								
CRI (Student-reported probability that teacher frequenty used CRI)	0.01	0.00	1.597		0.00	0.00	-0.354	
Pretests (ES and MS)								
Engagement, 7th grade	0.50	0.02	21.976	***	0.51	0.02	21.764	***
Engagement 7th grade-Squared	0.05	0.01	4.038	***	0.06	0.01	3.988	***
Engagement, 6th grade	0.17	0.02	7.124	***	0.18	0.03	6.897	***
Engagement 6th grade-Squared	-0.02	0.02	-1.211		-0.02	0.02	-0.881	
Math EOG 5th grade	0.00	0.00	-0.240		0.00	0.00	-0.259	
Reading EOG 5th grade	0.00	0.00	0.156		0.00	0.00	0.063	
Math EOG 4th grade	0.00	0.00	0.174		0.00	0.00	0.108	
Reading EOG 4th grade	0.00	0.00	0.680		0.00	0.00	0.857	
Math EOG 3rd grade	0.01	0.01	0.832		0.01	0.01	1.121	
Reading EOG 3rd grade	0.00	0.00	-0.940		0.00	0.00	-1.108	
Demographics and Family Variables								
Age	0.03	0.04	0.678		0.02	0.04	0.566	

Age-Squared	0.07	0.06	1.174		0.06	0.06	0.914	
Age-Cubed	-0.06	0.04	-1.356		-0.04	0.04	-1.026	
Black	0.13	0.04	3.550	***	0.12	0.04	3.135	**
Hispanic	0.05	0.05	1.123		0.04	0.05	0.734	
Other R/E (Amer Ind/Asian/Aleutian/Pac Isle/Multi/Other)	0.01	0.06	0.154		-0.01	0.06	-0.223	
Gender (female)	0.03	0.03	1.309		0.05	0.03	1.576	
Receives free/reduced price lunches	-0.12	0.04	-3.545	***	-0.11	0.04	-2.912	**
Academically gifted	0.01	0.04	0.312		0.00	0.05	0.026	
Enrolled in special education services	-0.09	0.05	-1.961	*	-0.09	0.05	-1.869	
Parent marital status (single)	-0.06	0.03	-1.950		-0.06	0.03	-1.870	
Teacher and School Characteristics								
SS teacher years of service					-0.01	0.00	-2.413	*
SS teacher gender: female					-0.01	0.04	-0.235	
SS teacher race: Black					0.31	0.07	4.324	***
Avg 7th grade math score of students with same 8th grade math teacher	0.01	0.00	1.207		0.00	0.00	1.078	
School average free/reduced lunch	-0.15	0.12	-1.231		-0.18	0.11	-1.557	

*** p < .001; ** p < .01; * p < .05;
[1]Number of teachers: math, 48; language arts, 45; science, 50; social studies, 42;
[2]Number of teachers: math, 33; language arts, 34; science, 33; social studies, 36.

Table 2.19.

Valuing, Math CRI IVE/LATE

	Main: All Individual Covariates, Teacher Assignment, School Level Controls[1]				Teacher Covs: All Individual Covariates, Teacher Assignment, School Level Controls; with Teacher Covariates[2]			
Variable	**Est.**	**SE**	**t value**		**Est.**	**SE**	**t value**	
Intercept (conditional mean for reference conditions)	1.83	2.51	0.729		2.09	3.08	0.678	
Independent Variable/Treatment Effect								
CRI (Student-reported probability that teacher frequenty used CRI)	0.00	0.00	1.860		0.01	0.00	1.754	
Pretests (ES and MS)								
Valuing, 7th grade	0.44	0.02	21.029	***	0.43	0.02	19.056	***
Valuing 7th grade-Squared	0.03	0.02	2.031	*	0.03	0.02	1.613	
Valuing, 6th grade	0.17	0.02	7.160	***	0.16	0.03	6.034	***
Valuing 6th grade-Squared	0.01	0.02	0.830		0.01	0.02	0.548	
Math EOG 5th grade	0.00	0.00	-0.239		0.00	0.00	0.081	
Reading EOG 5th grade	0.00	0.00	0.426		0.00	0.00	0.422	
Math EOG 4th grade	0.00	0.00	0.446		0.00	0.00	0.071	
Reading EOG 4th grade	0.00	0.00	0.900		0.00	0.00	0.336	
Math EOG 3rd grade	0.00	0.01	-0.455		0.00	0.01	0.061	
Reading EOG 3rd grade	0.00	0.00	-0.081		0.00	0.00	0.045	
Demographics and Family Variables								
Age	0.02	0.03	0.503		0.01	0.04	0.260	

Age-Squared	-0.01	0.05	-0.281		0.00	0.05	-0.031	
Age-Cubed	0.00	0.03	-0.110		0.00	0.04	0.093	
Black	0.19	0.03	6.418	***	0.19	0.03	5.712	***
Hispanic	0.11	0.04	2.680	**	0.10	0.04	2.222	*
Other R/E (Amer Ind/Asian/Aleutian/Pac Isle/Multi/Other)	0.14	0.05	3.085	**	0.16	0.06	2.933	**
Gender (female)	0.11	0.02	4.992	***	0.13	0.03	4.987	***
Receives free/reduced price lunches	-0.06	0.03	-2.123	*	-0.04	0.03	-1.278	
Academically gifted	-0.07	0.04	-1.855		-0.08	0.04	-1.917	
Enrolled in special education services	0.04	0.04	1.000		0.03	0.04	0.757	
Parent marital status (single)	0.01	0.02	0.537		0.01	0.03	0.486	
Teacher and School Characteristics								
Math teacher years of service					0.00	0.01	-0.010	
Math teacher gender: female					0.08	0.10	0.811	
Math teacher race: Black					0.02	0.10	0.199	
Avg 7th grade math score of students with this 8th grade math teacher	0.01	0.01	0.814		0.00	0.01	0.566	
School average free/reduced lunch	0.25	0.18	1.373		0.13	0.22	0.578	

*** p < .001; ** p < .01; * p < .05;

[1] Number of teachers: math, 48; language arts, 45; science, 50; social studies, 42;

[2]Number of teachers: math, 33; language arts, 34; science, 33; social studies, 36.

Table 2.20.

Valuing, LA CRI IVE/LATE

	Main: All Individual Covariates, Teacher Assignment, School Level Controls[1]				Teacher Covs: All Individual Covariates, Teacher Assignment, School Level Controls; with Teacher Covariates[2]			
Variable	**Est.**	**SE**	**t value**		**Est.**	**SE**	**t value**	
Intercept (conditional mean for reference conditions)	-1.15	2.30	-0.502		-1.06	3.82	-0.277	
Independent Variable/Treatment Effect								
CRI (Student-reported probability that teacher frequenty used CRI)	0.007	0.00	3.042	**	0.01	0.00	1.815	
Pretests (ES and MS)								
Valuing, 7th grade	0.44	0.02	21.134	***	0.44	0.02	19.263	***
Valuing 7th grade-Squared	0.03	0.01	2.098	*	0.03	0.02	1.911	
Valuing, 6th grade	0.16	0.02	6.912	***	0.16	0.03	6.082	***
Valuing 6th grade-Squared	0.01	0.02	0.510		0.00	0.02	-0.061	
Math EOG 5th grade	0.00	0.00	-0.259		0.00	0.00	-0.164	
Reading EOG 5th grade	0.00	0.00	0.371		0.00	0.00	0.583	
Math EOG 4th grade	0.00	0.00	0.478		0.00	0.00	0.028	
Reading EOG 4th grade	0.00	0.00	0.925		0.00	0.00	0.723	
Math EOG 3rd grade	0.00	0.01	-0.426		0.00	0.01	-0.322	
Reading EOG 3rd grade	0.00	0.00	-0.016		0.00	0.00	0.033	
Demographics and Family Variables								
Age	0.01	0.03	0.403		0.02	0.04	0.575	

Age-Squared	-0.01	0.05	-0.224		-0.01	0.05	-0.236	
Age-Cubed	0.00	0.03	-0.075		-0.01	0.04	-0.347	
Black	0.19	0.03	6.476	***	0.20	0.03	5.815	***
Hispanic	0.11	0.04	2.737	**	0.12	0.04	2.735	**
Other R/E (Amer Ind/Asian/Aleutian/Pac Isle/Multi/Other)	0.14	0.05	3.056	**	0.15	0.05	2.870	**
Gender (female)	0.11	0.02	4.928	***	0.12	0.03	4.576	***
Receives free/reduced price lunches	-0.06	0.03	-2.035	*	-0.07	0.03	-2.197	*
Academically gifted	-0.08	0.04	-2.231	*	-0.08	0.04	-2.044	*
Enrolled in special education services	0.06	0.04	1.407		0.05	0.05	1.130	
Parent marital status (single)	0.01	0.02	0.615		0.02	0.03	0.773	
Teacher and School Characteristics								
LA teacher years of service					0.00	0.00	1.048	
LA teacher gender: female					0.01	0.07	0.132	
LA teacher race: Black					0.06	0.07	0.877	
Avg 7th grade reading score of students with this 8th grade LA teacher	0.01	0.01	2.196	*	0.01	0.01	1.273	
School average free/reduced lunch	0.32	0.14	2.360	*	0.34	0.18	1.913	

*** p < .001; ** p < .01; * p < .05;
[1]Number of teachers: math, 48; language arts, 45; science, 50; social studies, 42;
[2]Number of teachers: math, 33; language arts, 34; science, 33; social studies, 36.

Table 2.21.

Valuing, Science CRI IVE/LATE

	Main: All Individual Covariates, Teacher Assignment, School Level Controls[1]				Teacher Covs: All Individual Covariates, Teacher Assignment, School Level Controls; with Teacher Covariates[2]			
Variable	**Est.**	**SE**	**t value**		**Est.**	**SE**	**t value**	
Intercept (conditional mean for reference conditions)	3.07	1.41	2.173	*	4.85	1.58	3.072	**
Independent Variable/Treatment Effect								
CRI (Student-reported probability that teacher frequenty used CRI)	0.007	0.00	3.454	***	0.008	0.03	0.257	
Pretests (ES and MS)								
Valuing, 7th grade	0.44	0.02	21.111	***	0.43	0.02	18.322	***
Valuing 7th grade-Squared	0.03	0.01	2.023	*	0.03	0.02	1.590	
Valuing, 6th grade	0.17	0.02	7.030	***	0.17	0.03	6.418	***
Valuing 6th grade-Squared	0.01	0.02	0.632		-0.02	0.02	-0.946	
Math EOG 5th grade	0.00	0.00	-0.085		0.00	0.00	-0.159	
Reading EOG 5th grade	0.00	0.00	0.282		0.00	0.00	0.067	
Math EOG 4th grade	0.00	0.00	0.526		0.00	0.00	0.094	
Reading EOG 4th grade	0.00	0.00	0.855		0.00	0.00	0.938	
Math EOG 3rd grade	0.00	0.01	-0.332		0.00	0.01	-0.055	
Reading EOG 3rd grade	0.00	0.00	-0.134		0.00	0.00	0.120	
Demographics and Family Variables								
Age	0.02	0.03	0.494		0.01	0.03	0.401	

Age-Squared	-0.01	0.05	-0.143		-0.01	0.05	-0.233	
Age-Cubed	-0.01	0.03	-0.168		0.00	0.04	0.035	
Black	0.20	0.03	6.548	***	0.20	0.03	6.052	***
Hispanic	0.10	0.04	2.656	**	0.09	0.04	2.103	*
Other R/E (Amer Ind/Asian/Aleutian/Pac Isle/Multi/Other)	0.14	0.05	2.975	**	0.15	0.05	3.041	**
Gender (female)	0.11	0.02	4.843	***	0.10	0.02	3.881	***
Receives free/reduced price lunches	-0.06	0.03	-1.957		-0.07	0.03	-2.240	*
Academically gifted	-0.08	0.04	-2.318	*	-0.07	0.04	-1.893	
Enrolled in special education services	0.05	0.04	1.369		0.04	0.05	0.798	
Parent marital status (single)	0.01	0.02	0.470		0.02	0.03	0.638	
Teacher and School Characteristics								
Science teacher years of service					0.00	0.03	-0.113	
Science teacher gender: female					0.00	0.55	0.009	
Science teacher race: Black					0.01	0.76	0.016	
Avg 7th grade math score of students with same 8th grade math teacher	0.00	0.00	0.558		0.00	0.00	-0.641	
School average free/reduced lunch	0.19	0.09	2.287	*	0.04	1.08	0.038	

*** p < .001; ** p < .01; * p < .05;
[1]Number of teachers: math, 48; language arts, 45; science, 50; social studies, 42;
[2]Number of teachers: math, 33; language arts, 34; science, 33; social studies, 36.

Table 2.22.

Valuing, SS CRI IVE/LATE

	Main: All Individual Covariates, Teacher Assignment, School Level Controls[1]				Teacher Covs: All Individual Covariates, Teacher Assignment, School Level Controls; with Teacher Covariates[2]			
Variable	**Est.**	**SE**	**t value**		**Est.**	**SE**	**t value**	
Intercept (conditional mean for reference conditions)	4.32	1.42	3.045	**	3.40	1.53	2.222	*
Independent Variable/Treatment Effect								
CRI (Student-reported probability that teacher frequenty used CRI)	0.01	0.00	2.925	**	0.01	0.00	2.301	*
Pretests (ES and MS)								
Valuing, 7th grade	0.44	0.02	21.628	***	0.44	0.02	19.549	***
Valuing 7th grade-Squared	0.03	0.01	2.325	*	0.02	0.02	1.333	
Valuing, 6th grade	0.16	0.02	6.854	***	0.17	0.03	6.849	***
Valuing 6th grade-Squared	0.01	0.02	0.612		0.02	0.02	0.830	
Math EOG 5th grade	0.00	0.00	-0.128		0.00	0.00	0.137	
Reading EOG 5th grade	0.00	0.00	0.288		0.00	0.00	-0.306	
Math EOG 4th grade	0.00	0.00	0.518		0.00	0.00	0.246	
Reading EOG 4th grade	0.00	0.00	0.874		0.00	0.00	1.078	
Math EOG 3rd grade	0.00	0.01	-0.378		0.00	0.01	-0.049	
Reading EOG 3rd grade	0.00	0.00	-0.081		0.00	0.00	-0.342	
Demographics and Family Variables								
Age	0.01	0.03	0.449		0.01	0.03	0.170	

Age-Squared	-0.01	0.05	-0.205		-0.01	0.05	-0.262	
Age-Cubed	-0.01	0.03	-0.168		0.00	0.03	0.047	
Black	0.20	0.03	6.575	***	0.20	0.03	6.201	***
Hispanic	0.11	0.04	2.777	**	0.11	0.04	2.709	**
Other R/E (Amer Ind/Asian/Aleutian/Pac Isle/Multi/Other)	0.14	0.05	3.050	**	0.13	0.05	2.704	**
Gender (female)	0.11	0.02	5.046	***	0.12	0.02	4.753	***
Receives free/reduced price lunches	-0.06	0.03	-2.002	*	-0.07	0.03	-2.142	*
Academically gifted	-0.07	0.04	-2.055	*	-0.07	0.04	-1.612	
Enrolled in special education services	0.06	0.04	1.416		0.06	0.04	1.345	
Parent marital status (single)	0.01	0.02	0.603		0.02	0.03	0.589	
Teacher and School Characteristics								
SS teacher years of service					0.00	0.00	-1.774	
SS teacher gender: female					0.01	0.03	0.427	
SS teacher race: Black					0.18	0.07	2.792	**
Avg 7th grade math score of students with same 8th grade math teacher	0.00	0.00	-0.295		0.00	0.00	0.363	
School average free/reduced lunch	0.03	0.11	0.304		0.04	0.10	0.377	

*** p < .001; ** p < .01; * p < .05;
[1]Number of teachers: math, 48; language arts, 45; science, 50; social studies, 42;
[2]Number of teachers: math, 33; language arts, 34; science, 33; social studies, 36.

Table 2.23.

Models for Academic Press and Preparation for Future Schooling as Endogenous Variable (Replaces CRI)

		Uncertain as Missing (and Imputed)	
Dependent Variable	**CRI Subject**	**Academic Press**	**Preparation for Future Schooling**
8th Grade Math	Math		
8th Grade Reading	LA		
Engagement	Math		.029*
	LA		
	Science		.031*
	Social Studies		.016*
Valuing	Math	.024*	.042***
	LA	.032***	.036***
	Science	.031***	.04***
	Social Studies	.029***	.034***

Models contain all possible student, teacher and school covariates.

CHAPTER 3. A NON-RANDOMIZED EFFICACY TRIAL REPLICATION OF A TEACHING PRACTICE INTERVENTION USING PROPENSITY SCORE WEIGHTING

Motivation is a cognitive or affective process that is related to student engagement, academic achievement, and learning (Wentzel & Wigfield, 2007). Interventions focused on motivation have the potential to improve student learning outcomes, and evidence-based teaching practice interventions that rely on motivational techniques are seen as one potential policy instrument for improving student performance (Roeser, Eccles & Sameroff, 2000; Wentzel & Wigfield, 2007). One such teaching practice intervention, CareerStart, supports teachers in their efforts to motivate students by associating lesson content with the knowledge and skills needed to be successful in a range of jobs and careers (Rose, Woolley, Orther, Akos, & Jones-Sanpei, 2012). A randomized control trial (RCT) demonstrated significant effects for CareerStart on school valuing, a measure of emotional engagement with schooling (Orthner, Jones-Sanpei, Akos & Rose, 2012), and achievement in middle school (Woolley, Rose, Orthner, Akos & Jones-Sanpei, in press) and high school (Woolley et al., revise & resubmit). This RCT was an efficacy trial conducted under idealized conditions including students and teachers from a single district and a partnership between researchers and district officials in ensuring a strong implementation. Given its success, an effectiveness trial, conducted under more realistic conditions including a more heterogeneous population and implementation that is independent of the developers and investigators (Fraser et al., 2009; Lee & Krajcik, 2012) should be conducted.

Despite the favorable evidence, important questions remain before scaling up to an effectiveness trial. First, the findings did not support a significant treatment effect of CareerStart on students' reading achievement. This may have been due to the higher variability between schools in reading than in math achievement. Whether CareerStart should continue to be implemented in language arts would be valuable information to have in re-developing CareerStart for an effectiveness trial. Second, it would be useful to understand the role that district support played in the efficacy trial, as this support is not likely to be replicated elsewhere. A non-random dissemination of CareerStart has presented a unique opportunity to address both of these needs. Thirteen schools outside of the efficacy trial implemented CareerStart non-randomly beginning in 2007-2008. These schools had a more heterogeneous population than the original RCT. Further, on average, they received less contact and fewer site visits from the person responsible for implementation of CareerStart, the CareerStart Coordinator, than did the original RCT schools. Strategic use of this non-random implementation, which also included the 7 control schools from the RCT, provided additional information that can be used to refine CareerStart for evaluation in an effectiveness trial.

In this study, I use quasi-experimental methods to compare students in the 20 schools non-randomly implementing CareerStart—including 13 outside the efficacy trial district and 7 within this district—with students from other middle schools in North Carolina to test the hypothesis that CareerStart promoted math and reading achievement. I use propensity score analysis (PSA) methods, with inverse probability weighting (Hirano, Imbens & Ridder, 2003), to model an interrupted time series of student performance from third to eighth grade, replicating the analysis design in Woolley et al. (in press). This design was supplemented by

additional parameters intended to answer the question of whether the strength of the implementation in the seven former RCT controls was greater relative to the thirteen non-RCT schools, and to establish the credibility of the causal inference of the PSA design, which is typically based on the strong assumption that observed variables sufficiently control for the influence of unobserved covariates of treatment and the outcome (Rosenbaum, 2002). Because the seven RCT treatment schools continued to implement CareerStart during the non-random study period, evidence showing that the PSA and RCT estimates for these seven schools are statistically equivalent may support the credibility of the PSA design in controlling for the influence of unobserved confounders. A future effectiveness study of CareerStart would benefit from answers to these questions.

In the next sections, I describe CareerStart, prefacing this description with a review of engagement and motivation, and relevance as a teaching practice to promote these student processes. This is followed by a summary of evidence from the RCT. I then review the importance of replication and scalability of effective programs, describing extant research on replication of efficacious motivation programs. I provide evidence that CareerStart may meet the conditions for examination in a future effectiveness trial, but may have unanswered questions that a PS-based design applied to the existing non-random replication sample can help answer. I describe the Rubin Causal Model (RCM), which explicates requirements that study parameters must meet to be interpretable as causal, and detail the challenges and opportunities associated with using PSA designs for replication. I then discuss the methods and results of this PSA study, followed by policy and research implications.

CareerStart: Engagement, Motivation and Relevance

Motivation is defined as "the process by which goal-directed activity is instigated and sustained" (Schunk, Pintrich, & Meece, 2008, p. 4)" or a "set of processes that provides energy for different behaviors" (Wentzel & Wigfield, 2007, p. 262). Two historical tendencies in instructional practice present significant challenges for motivating students to learn, and present opportunities for relevance-oriented teaching practices. First, teachers present new material deliberately in highly decontextualized forms to ensure it has the highest generalizability (Cordova & Lepper, 1996; Dewey, 1916). In doing so, what makes the material interesting to students may be lost (Cordova & Lepper, 1996). Second, school is by its very nature future-oriented. This suggests two additional problems. If the future is not seen as positive and hopeful, it may be an insufficient motivation (Phalet, Andriessen & Lens, 2004). Further, if the future goals cannot be translated into more proximal near-term goals, then they may seem too far out of reach to be realistic motivators (Husman & Lens, 1999). Content can be imbued with relevance by linking the content to concepts that are relevant to students' present or future lives (Cordova & Lepper, 1996; Husman & Lens, 1999). Several theories suggest processes that transmit the benefit of relevance on engagement through student motivation. These include expectancy value theory (Eccles & Wigfield, 2002); interest theories (Hidi & Harackiewicz, 2000; Schiefele, 1991); goal orientation theories (Ames, 1992; Harackiewicz, Barron, Pintrich, Elliot & Thrash, 2002); self-determination theory (Lepper & Cordova, 1992; Ryan & Deci, 2000); and identity theory (Markus & Nurius, 1986).

Relevance to students' current lives has been shown to be an effective motivator (Cordova & Lepper, 1996). Geier et al. (2008), for example, demonstrated long-term learning

effects from a science-based inquiry curriculum that included content relevant to students' lives. Given the future orientation of schooling, however, practices that link relevance for the future to present learning may help students realize the value of present-day competence at content and tasks (Husman & Lens, 1999). For example, a study with young children demonstrated that when future careers were seen as education dependent, students invested more time and effort in schoolwork (Destin & Oyserman, 2010). Oyserman, Terry and Bybee (2002) showed that a program that helps youth focus on or imagine themselves as successful adults, and then related these images to current school activity motivation, was associated with better behaviors and engagement.

One type of relevance to students' future lives that has demonstrated promise as a means of promoting intrinsic motivation through interest, goals and autonomy processes is career relevance. Studies show that even middle school students consider college and career plans and identities (Akos, Konold & Niles, 2004), and must make curricular choices that will in part determine future educational opportunities (Akos & Galassi, 2004). By tying tasks and material to specific jobs and careers, students may make better choices regarding engagement in school, courses taken, and completion of schoolwork (Husman, Derryberry, Crowson & Lomax, 2004; Husman & Lens, 1999).

CareerStart

CareerStart is a teaching practice intervention that improves the relevance of lesson content by associating the skills and knowledge taught with information from careers and jobs that use these skills. CareerStart provides direct and indirect assistance to teachers to help teachers convey to students information about the value of the material being taught, which I labeled career relevant instruction (CRI; Rose et al., 2012). This value is expressed

as a distal goal in the form of competency in tasks that may have some utility later in life during a job or career. One lesson, for example, uses hurricane disaster preparation as a context to teach science; another uses messages in advertising to teach language arts (Woolley et al., in press). Tasks that are presented as useful to students' career identities, or career-relevant, during this formative period may help students translate distal career goals and aspirations into current learning goals (Husman & Lens, 1999). This may alter student's affect towards and enhance the authenticity of the material and tasks, increasing students' motivation to learn (Husman, Derryberry, Crowson, & Lomax, 2004). CareerStart consisted of seven interwoven components.

Coordinator. A CareerStart Coordinator was responsible to schools for organizing and implementing CareerStart. The Coordinator had three primary responsibilities. The first was to discharge formal implementation components including organizing and leading trainings and quarterly meetings, acting as a resource to the teachers, for example by responding to questions about lesson modification, and preparing and distributing a newsletter to teachers. A second responsibility was to serve as the focal point for the informal distribution of information between schools. In this regard, the Coordinator was at the center of a "web" of information about CareerStart spanning the schools implementing the program, and teachers used this resource to share information between schools (about, for example, the modification of lessons and development of lesson plans). Third, the Coordinator acted as a liaison between the schools and the evaluation team, providing information to the evaluation team about progress of implementation and sharing data obtained from the evaluation team with teachers and principals. The coordinator was central to implementation of many of the other components.

Staggered grade-level implementation & training. Program fidelity called for CareerStart to be implemented one grade level at a time, starting with sixth grade and progressing through eighth grade. The first year of implementation was considered a training year in which teachers would learn on the job. This was a treated year, but the cohort of students from this year was not evaluated. Evaluation began in the second year of implementation. It was expected that a high fidelity implementation would be sustained after the training year in each grade level.

Packaged lessons. The main implement of CareerStart was a set of 10 pre-packaged lessons in each of the four core subject areas—math, language arts, science and social studies—for a total of 40 lessons per grade level over all three middle grade levels, plus 10 extra lessons for students taking algebra in eighth grade. These lessons were developed by teachers in the RCT district with the aid of the Coordinator, and reviewed by curriculum specialists to ensure that they met the requirements of the standard course of study. These lessons associated the tasks and materials to be learned with specific jobs and careers. Implementation fidelity called for all 10 lessons to be taught, though the lessons could be modified as needed to fit students' needs and interests and teacher style.

Teacher and curriculum coordinator buy-in. Principals were discouraged from simply pressuring their faculty to implement the program, but rather were encouraged to sell the program to the faculty. Teachers and curriculum coordinators who demonstrated enthusiasm for the program were encouraged to sell the program to their colleagues at their own school and at other treatment schools, as it was theorized that teachers would trust the views of other teachers. Enthusiastic support for CareerStart among faculty was considered a key element toward implementation fidelity. Buy-in was intended to address concerns that

the program would be seen as something important to the superintendent that was then subsequently imposed on the teachers. By helping teachers to understand how students could benefit and getting enthusiastic support from them, it was proposed that the teachers would take ownership of the program. In the efficacy trial, support by faculty was not measured directly, but could be measured indirectly using questions on a teacher survey regarding the importance of career relevance to student learning.

Training. A training session was held at the start of every school year with teachers and principals new to CareerStart. The training consisted primarily of a slideshow and handouts that focused on the benefits of relevance to student learning. As data were collected during the evaluation, the trainings were supplemented with these data to demonstrate that the program was potentially valuable. Having a training session at each school at the beginning of the year was a central part of implementation fidelity, and qualitative data were collected to document fidelity in this area. Principal meetings were "joint" spanning multiple schools to encourage collaboration across schools.

Supplementation. Teachers were encouraged to consider the career relevance of the other content they teach and develop career relevant lessons on their own or in teacher teams. Additional supports to teachers were provided in the form of a career newsletter to help them develop additional career relevant lessons, and schools were encouraged to hold career fairs that would encourage career interest among the students as well as help faculty understand the links between content they were teaching and jobs and careers in the local labor market. Supplementation was a key part of implementation, but not measured directly. Information about the newsletters that were distributed and the career fairs held at each school are

available, and qualitative data suggested that teacher teams worked together at some schools to develop supplemental lessons.

Planning meetings. Quarterly planning meetings were to be held at each school for the faculty to develop lessons plans that incorporated the pre-packaged lessons as well as supplemental lessons. The Coordinator organized and led these meetings, and used these meetings as opportunities to distribute information about activities taking place at other schools. Teachers also completed accountability forms during these meetings. High fidelity implementation called for having a quarterly meeting in each school. Interview data provide limited documentation of fidelity in this area.

Evidence of the Efficacy of CareerStart

A randomized efficacy trial was conducted from 2004-05 to 2008-09 in the North Carolina school district in which CareerStart was developed. In this design, the 14 middle schools in the district at the time were randomly assigned to two conditions of seven schools each. During this evaluation, key portions of the CareerStart logic model been validated. For teachers, CareerStart has been demonstrated as promoting the use of CRI among teachers in math (odds = 1.879, $p < .001$), including a stronger effect for teachers with more years of service (at 1 standard deviation above the mean, odds = 2.171; Rose et al., 2012).

Two qualitative investigations using focus group data obtained from teachers and students lent support to the findings of Rose et al. (2012). A fidelity study showed that in three of the seven treatment schools (labeled “low fidelity”), teachers used well less than half of the lessons on average and that in contrast to teachers in a selected high fidelity school, who reported modifying the lessons to fit their students’ needs, teachers in a low fidelity school elected not to use the lesson when they perceived the fit to be poor (Phillippo & Rose,

2012). In a second qualitative study teachers expressed concern about the fit of the lessons to their students' needs (Woolley, Rose, Mercado & Orthner, 2013).

For students, CareerStart demonstrated a causal effect on students' valuing of school ($b = 1.41$, $p < .01$; Orthner et al., 2012). In Orthner et al. (2012), valuing ($b = 1.31$, $p < .01$) and engagement ($b = 1.59$, $p < .001$) were also demonstrated to be associated with the level of CRI used by teachers. The qualitative study lent support to these findings; some teachers reported that the lessons increased students' involvement in the classroom and promoted engaging classroom discussions (Woolley et al., 2013).

Woolley et al. (in press) demonstrated that CareerStart affected the ultimate outcome of student achievement. CareerStart was demonstrated to have significantly increased students' end-of-grade math achievement in middle school, at 0.61 points on the EOG test scale; this is equivalent to one quarter of a standard deviation effect size and about 1/3 the size of the annual rate of growth in math for the average student. The math effect size of 0.25 standard deviations was found to be significant at $p < .05$, but the much larger reading effect at 0.46 standard deviations was not (Woolley et al., in press). The between-school variance in the reading outcome inflated the standard error for the treatment effect in reading relative to that for math. The study may have been underpowered for detecting effects that are typical of educational interventions.

Finally, CareerStart had lasting effects on children in high school at the time these students were in ninth grade, with students from CareerStart middle schools performing better on end-of-course biology tests ($b = 2.52$, $p < .001$) and accruing credits towards graduation ($b = .21$, $p < .001$) at a higher rate than students from control middle schools (Woolley et al., revise & resubmit).

The sum of this evidence suggests that CareerStart is a promising and potentially evidence-based program. However, these findings were obtained under idealized conditions typical of efficacy studies. Evidence-based support will require an effectiveness study of CareerStart. In the next section, I describe scalability and compare efficacy and effectiveness designs, the requirements that studies must typically meet to be candidates for effectiveness investigations, and then explain how the evidence supports that CareerStart satisfies these conditions.

Intervention Scalability

A stated goal of research into educational and curriculum innovation is to develop and refine programs that can be disseminated, potentially as systemic reform (Penuel, Fishman, Cheng & Sabelli, 2011). A typical education program, if brought "to scale," would go through three stages: efficacy, effectiveness and dissemination (McDonald, Keesler, Kauffman & Schneider, 2006). Blumenfeld et al. (2000) defined a scalable program as one that could be implemented outside of the resource-intensive environment of an efficacy or effectiveness trial, suggesting that *scale* is synonymous with dissemination. Another definition suggests that scalability research involves an increase in the number of contexts of implementation (Fuchs & Fuchs, 1998). According to Clarke and Dede (2009) and Coburn (2003), scale encompasses multiple normative dimensions including the depth of changes, sustainability of these changes over long duration, diffusion to multiple heterogeneous settings, shift in ownership of the program to its users, and evolution of the program to fit local needs. Scalability, further, may imply *transportability* of effective programs to settings in which they were not specifically tested (Schoenwald & Hoagwood, 2001).

Scaling from Efficacy to Effectiveness

Distinct phases of evaluation are critical to the cost-effective development of policy interventions including instructional practice interventions such as CareerStart, and only programs surpassing certain standards of evidence at one stage should be admitted to the next, more costly, stage. For example, the Institute for Education Sciences (IES) specifies efficacy and effectiveness as distinct goals for researchers seeking funding for their projects (2012). Generally, efficacy studies are intended to examine the causality of a program in promoting its intended outcomes and to rule out plausible alternative explanations for treatment effects (Lee & Krajcik, 2012), but are typified by settings that are unlikely to be realized in a real-world setting (Slavin, 2008). These may include a homogeneous population of students and teachers from a cooperating school or district (Lee & Krajcik, 2012) and a higher level of support for implementation from the investigator than would be expected in a real-world setting (Fraser et al., 2009). In the case of CareerStart, this cooperation went as far as to include having teachers in the district develop the lessons and to have frequent sharing of outcome data with teachers.

An effectiveness study, alternatively, requires that the program be studied under more realistic conditions, including a more heterogeneous population and implementation that is independent of the developers and investigators as it would be as a systemic reform (Lee & Krajcik, 2012). In taking a program from efficacy to effectiveness evaluation, the role of the idealized context of the efficacy trial should be examined carefully (Fixsen, Blasé, Naoom & Wallace, 2009; Schoenwald & Hoagwood, 2001). Some compromises may be needed between fidelity of the intervention as described and demonstrated as efficacious in the

smaller context, and the constraints imposed by the larger more heterogeneous context (Lee & Krajcik, 2012).

O'Donnell (2008) suggested that two dimensions—program fidelity and positive outcomes—inform the decision to scale up an intervention, and recommends that only programs demonstrating both effectiveness and fidelity should be scaled up. Elias, Zins, Graczyk, and Weissberg (2003) described several implementation factors that support sustainability at scale of evidence-based educational programs, including requirements for personnel at the school and district levels, training, and visibility of the program in the school and community. Flay and colleagues (2005) proposed five specific criteria for the level of evidence needed for a *prevention* program to scale up from efficacy to effectiveness: a) the evaluation must be conducted on a defined sample from a defined population that is the target of the program; b) measures and data collection techniques must be scientifically sound; c) studies must have sufficient analytic rigor; d) consistent positive treatment effects and no adverse effects from the treatment on either the treatment or control groups; e) at a minimum at least one program effect must be demonstrated over the long-term. Although these recommendations were not designed for educational programs, they are highly specific criteria that may be informative for education evaluators to consider.

Extant Research on the Scalability of Motivational Instructional Practices

Research on taking motivational instructional interventions to scale is limited. However, the few studies that are available are instructive in that they provide a clearer picture of the challenges described above as they apply to scalability of CareerStart. Bishop, Berryman, Wearmouth, Peter, and Clapham (2012) described a non-randomized within-study replication of an evaluation of a teaching practice intervention for improving the motivation

and engagement of indigenous students in New Zealand schools. The authors looked at two phases of implementation of this program (labeled phase 3 and phase 4), that took place in consecutive periods. Phase 4 had a larger number of settings (n = 21) than did phase 3 (n = 12). The program targeted change in teachers' interactions with students (e.g., demonstrating higher regard for student background, and modifying their instructional methods accordingly) and consisted of supports that the authors acknowledged could not be sustained. Despite the larger group of implementation sites, with the level of support maintained at high levels, phase 4 schools replicated the results of the phase 3 schools.

Lynch, Pyke and Grafton (2012) examined implementation and scale-up of three middle school science curriculums. The authors showed that effects observed at the efficacy scale were not replicated on a larger scale; the authors subsequently used qualitative methods to examine the context of the scale-up to the effectiveness study to identify factors that may have impeded successful implementation at larger scale. Among other findings, the authors indicated that administrative, student and teacher support may have been inadequate. Blumenfeld et al. (2000) also used qualitative methods to study the fit between school capacity, policy, culture and the intervention as designed, and discussed the challenges in adapting an intervention across multiple contexts. The authors found that district and administrator support, including financial support, influenced effectiveness, that teacher beliefs may have impeded widespread adoption, and that teachers needed time to learn the basic approaches and to adapt the program to fit their own context. Buzhardt, Greenwood, Abbott and Tapia (2006), in a school-randomized trial examined barriers to the scale-up of an evidence-based peer tutoring program, identified failure to communicate with program staff as the strongest predictor of poor implementation.

Tatar and colleagues (2008) reported findings from a non-random study of 21 seventh grade math teachers in a wide variety of educational settings using a instruction practice intervention previously shown efficacious in another setting. The authors found that short-term programs that lacked extensive professional development requirements, but that provided training and materials to use in instruction, were effective. Such programs may be more agile in that they can make an impact more quickly than those that require long-term commitments and long periods in professional development training.

Taking CareerStart to Scale: An Effectiveness Trial

The sum of the evidence from the CareerStart efficacy trial and the extant research on the scalability of motivational interventions that I described above suggests that scalability of CareerStart may be realistic given the agile packaged components such as the lessons and brief trainings (e.g., Tatar et al., 2008); the importance of district and administrator support and teacher buy-in (Blumenfeld et al., 2000; Lynch, Pyke & Grafton, 2012); and the role of communication among stakeholders (Buzhardt, Greenwood, Abbott & Tapia, 2006). However, they also suggest several problems that CareerStart must overcome to have a successful effectiveness trial.

The evidence cited above, which indicates that CareerStart meets or exceeds the Flay et al. (2005) criteria for scalability to an effectiveness trial largely owing to a middle school math treatment effect and a high school science effect, suggests that a decision must be made about the role that language arts lessons will play in an effectiveness trial. The RCT may have demonstrated that career relevance does not work in language arts, but it is also plausible that it may have demonstrated that the study was underpowered or contaminated, or that it was simply typical of average educational interventions in not promoting achievement

in reading. First, the treatment was administered to entire schools, and with only 14 middle schools in the district at the outset of the study, only seven schools were randomized to each treatment condition. This type of study in which schools rather than students are assigned to treatment, known as a cluster randomized trial, is typically underpowered (Hedges & Hedberg, n.d.; Schochet, 2005). In the case of CareerStart, the small number of middle schools in the evaluation may suggest that the design was underpowered for detecting a significant treatment reading effect. The other possibility was control group contamination. Although no direct evidence of CareerStart program lesson use or other measures of fidelity in the control schools was available, Rose and colleagues (2012) established that control school teachers did use CRI. Although control school math teacher used significantly less CRI than their treatment school counterparts, the differences in language arts, science and social studies were not significant, suggesting that some level of contamination may have occurred. Finally, reading effects are often not seen in educational interventions, and one explanation that has been offered is that the material tested on the reading exam may not be linked as closely to instruction in language arts as tested math content is to math instruction (Harris & Sass, 2011).

In addition, however, CareerStart has not been extensively studied for the intensity of the support provided, largely in the form of the CareerStart Coordinator, or its fit to local conditions, factors noted by O'Donnell (2008) and Elias, Zins, Graczyk and Weissberg (2003) as important to understand before introducing an intervention to a new environment. As noted, Bishop and colleagues (2012) were able to replicate a treatment effect in a larger sample but only if a level of support that would be impractical for systemic reform was maintained. Buzhardt, Greenwood, Abbott & Tapia (2006) indicated that communication

among stakeholders was important for fidelity; and Tatar and colleaguegs (2008) demonstrated the importance of brief training sessions. As all of these features are part of the Coordinator's responsibilities, a key question regarding the long-term and widespread sustainability of CareerStart is whether the position of CareerStart Coordinator, which requires funding, is replicable in general settings. It was also not clear how many schools a single Coordinator could serve and maintain an effective treatment.

Finally, the local conditions for the efficacy trial have not been examined. As noted, a key feature of the program was buy-in from principals and teachers for the concepts and practices of CareerStart. This was intended to address concerns that teachers would see the program as imposed on them by an enthusiastic superintendent who was heavily involved in its development and dissemination to neighboring districts. This could bias the program effect upward if teachers perceived the superintendent's support positively. Alternatively, the superintendent's support could bias the program effect downward if they perceived the superintendent's support negatively. A negative perception might occur for example if teachers see the program only as something of importance to the superintendent without any commensurate value as a teaching practice. As a consequence, if buy-in had not been successfully applied in the RCT and program success could be attributed instead to pressure from the superintendent, there may have been "home-district" effects that would not be present and not generalizable to other settings.

Questions about the power for testing a reading effect, about the importance and role of the coordinator, and buy-in vs. atypical local conditions would usually only be answerable in a subsequent effectiveness trial, but an opportunity has become available to respond to these needs and provide more information to program developers prior to scaling up by

examining the effect of a non-random replication of CareerStart on math and reading achievement in schools in the RCT district and outside this district.

Non-Random Replication of CareerStart

During the period of the efficacy RCT, neighboring schools and districts in central North Carolina expressed an interest in implementing CareerStart and the program was disseminated over a period of three years (2007-08 to 2009-10) to 13 additional schools in 7 districts (see Table 3.1A and 3.1B). Also during this period, the 7 control schools in the RCT were admitted to the treatment condition as part of a 2-year graded waitlist design; *i.e.*, two years after first implementing at the sixth grade level in 2005-06, sixth grade teachers in the control schools implemented CareerStart in 2007-08 (followed by seventh and eighth the subsequent two years). After imposing the one-year training period, these schools became available for evaluation starting in 2008-09.

The 20 schools implementing CareerStart non-randomly represented an opportunity to provide key information to further develop CareerStart for the purpose of an effectiveness study. First, the sample of treatment schools is nearly three times the size of the original RCT treatment group. Second, the schools implementing the treatment were more heterogeneous, spanning seven districts and two regions of the state (Table 3.2). Students in the RCT district and outside the RCT district differed significantly on all pre-treatment outcomes except student receives free/reduced lunch, percent Black students, and teacher is full time or receives full time pay. Finally, the presence of both 13 new treatment sites and the 7 RCT control schools in treatment that started implementing at the same time presents an opportunity to test whether schools in the RCT district have greater effects due to a higher level of support. Three questions about implementation in the 20 non-randomly assigned

schools guide this study; first, whether the overall non-random treatment promoted higher math and reading achievement; second, whether there were differences between this treatment and the treatment administered during the RCT; and third, whether there were differences between schools in the RCT district and in other districts. A fourth question addresses the credibility of the impact estimates from PSA.

The central challenge of a non-randomized evaluation design is obtaining a comparison group that mimics a randomized control group in its equivalence (Morgan & Winship, 2007). In this study, I examine these 20 schools for a treatment effect using a replication of the interrupted time series design of the efficacy RCT (Woolley et al., in press), including imposing a one-year wait time on implementing schools prior to evaluating, such that teachers may learn the treatment. Thus, the sixth grade cohort would be obtained from most of these schools in the school year 2008-09. I use PSA to obtain a weighted comparison sample using the entire middle school population in North Carolina starting sixth grade in 2008-09. PSA is controversial, with some scholars suggesting that the method inadequately controls for unobservable characteristics (e.g., Wilde & Hollister, 2007), and other scholars suggesting that PSA, used in conjunction with other controls and designs, may be adequate (e.g., Cook, Shadish & Wong, 2008). Prior to discussing PSA in detail and discussing the controversy over their utility to educational evaluation, I describe a framework, the Rubin Causal Model (Holland, 1986), which explicates the challenges PSA must overcome.

Propensity Score Analysis and the Rubin Causal Model

The CareerStart efficacy study was conducted using an RCT. Randomization is widely seen as the gold standard for establishing the causal effects of treatments because randomization equalizes both observed and unobserved confounders of the treatment across

treatment groups. A non-random study must effectively replicate such conditions for inferences about the causality of observed treatment effects to be credible.

In the Rubin causal model (RCM; Holland, 1986), also known as the potential outcomes model, the definition of the causal estimand for a treatment such as a teaching practice intervention depends on the conditions that are experienced in the absence of the specified cause, known as the counterfactual. In the case of a teaching practice intervention such as CareerStart that is given to entire schools, this is typically the standard teaching practices that teachers in each school use. Formally, assume that the outcome for student i (with I = 1,…N) under treatment condition j is Y_{ij} with two possible treatments, j = 1= CareerStart school, and j = 0 = non-CareerStart school (Morgan & Winship, 2007). Second, the student has one potential outcome Y_{ij} under each possible treatment. At most one of these two potential outcomes can be realized, as each student can only be assigned to one condition (the fundamental problem of causal inference; Holland, 1986). To circumvent this, the treatment effect is defined as a function of the distributions of students assigned to treatment and the students under the control, usually labeled the average treatment effect (*ATE*).

There are three assumptions supporting the inference of a distributional function such as the *ATE* as the causal effect of the treatment for an individual student. Two of these cannot be proven but nevertheless are assumed to be present in both randomized and observational designs (Morgan & Winship, 2007): First, manipulability implies that students can be assigned to either condition, which is required for both potential outcomes to be defined. Second, the stable unit treatment value assumption, implies that the potential outcome under either condition is not affected by peers also assigned to that condition. A third assumption, unconfoundedness (also referred to as ignorability or exogeneity), is directly addressed by

randomization and must be supported in a non-randomized setting if such a setting is to be used for causal evidence. Unconfoundedness implies that each student's assignment (*A*) to treatment *j* is independent of his or her potential outcome in that condition (Morgan & Winship, 2007):

$$\text{(i) } Y_{ij} \perp \text{A}$$

Condition *i* is assumed to be met in randomized designs. In non-randomized designs, typically a weaker statement conditions on background factors, x:

$$\text{(ii) } Y_{ij} \perp \text{A} \mid \text{X} = \text{x}$$

In other words, the potential outcome Y_{ij} under each condition is independent of assignment to treatment conditional on a set of characteristics *X*. Condition *ii* is violated in the presence of association between unmeasured predictors of the outcome that are also associated with each student's assignment to treatment. In an RCT, the treatment effect can be viewed as a simple difference between the expectations or average effects of those in the treated and non-treated groups (Morgan & Winship, 2007), relying on condition *i*: ATE = E[d] = E[Y_{i1}] – E[Y_{i0}] = E[Y_{i1}– Y_{i0}] with 1 = treated, 0 = control. Non-random designs must effectively duplicate this using condition *ii*, as ATE = E[d] = E[Y_{i1}| X = x] – E[Y_{i0}| X = x] = E[Y_{i1}– Y_{i0}| X = x].

PSA may meet this criterion. The propensity score is the probability *P* of assignment to the treatment condition (j = 1) conditional on factors related to assignment (Rosenbaum, 2002; Rosenbaum & Rubin, 1983). This can be represented as $P = \Pr[Z_{j=1} \mid X]$, with Z representing a "switch" that assigns a participant to each condition. If the true propensity score is known then both observable and unobservable covariates are effectively randomized. Typically the propensity score is not known, but alternatively if all of the predictors of the

propensity score are known, measured, and included in the propensity score model then the treatment conditions can be considered conditionally random. In both cases, this would result in treatment assignment that is conditionally independent of covariates of the outcome, conditional on the propensity P, as follows:

$$\text{(iii) } Z_{ij} \perp \text{X} \mid P$$

As a consequence, E[d] can be interpreted as a causal effect (Rosenbaum, 2002). This places the burden for satisfying unconfoundedness on the richness of the data.

Propensity Score Modeling

Because the propensity score is not typically known, it must be estimated using a conditional model consisting of a binary dependent variable describing probability of treatment assignment, $P = \Pr(Z_{j=1} = \{0, 1\})$ and a linear combination of covariates of the assignment (Rosenbaum, 2002). This can be implemented as follows: the probability of treatment assignment is converted to log-odds, as $(e) = ln(P/(1 - P))$. $L(e)$ is then regressed on a vector of predictors X as follows:

$$(1) \qquad L(e) = X\beta$$

The predicted values of this regression are reverse-transformed to probabilities, which are labeled the propensity scores. As RCM condition *ii* makes clear, the covariates of the assignment outcome should include at a minimum all variables that are both predictors of assignment and predictors of the outcomes of the treatment. A number of approaches to covariate selection have been proposed (Guo & Fraser, 2009; Imbens & Wooldridge, 2009). I discuss the approach used in this study in the methods.

This propensity score is then used to equalize participants in the treatment conditions, typically by matching or weighting. Participants in each condition can be matched on this

propensity using a number of different algorithms for matching (see, for example, Stuart & Green, 2008). Using the matching approach, participants in each condition that do not sufficiently overlap with the other condition are discarded, and a method of determining which participants should be matched to each other has to be determined. Typically matching methods focus on matching comparisons to each treatment and not the other way around. This implies that matching may not enable estimation of the ATE, but instead enable estimation of the average treatment on the treated (ATT), the average effect on the subpopulation that is treated (Imbens & Wooldridge, 2009). The ATT is a comparison between potential outcomes only among those participants in the treatment condition (Harder, Stuart & Anthony, 2010). However, the ATE, which is a comparison between all treatment and comparison individuals on their potential outcomes, is typically the estimand of interest (Guo & Fraser, 2009; Imbens & Wooldridge, 2009).

An alternative to matching, which may enable estimation of the ATE in most settings, is to use inverse probability weighting (Harder, Stuart & Anthony, 2010; Hirano, Imbens & Ridder, 2003; Imbens &Wooldridge, 2009; Morgan & Winship, 2007), whereby participants in the treatment and comparison conditions are assigned a weight proportional to their probability of treatment assignment. Weighting methods are more efficient because typically no comparison or treatment participants are discarded. Although weights enable estimation of the ATE, a version of inverse probability weighting can alternatively estimate the ATT by constraining the treatment group weights to one (Guo & Fraser, 2009). Matching and weighting methods can also be combined, by sequentially matching followed by weighting (Henry, Gordon & Rickman, 2006) or by weighting in the common support, which potentially implies an ATT estimate (Harder, Stuart & Anthony, 2010), or by trimming

comparison cases from the lower unsupported range, leaving high probability treatment participants in the analysis to support an ATE estimate (Harder, Stuart & Anthony, 2010). Both matching and weighting methods can be combined with regression approaches (Glazerman, Levy & Myers, 2003; Imbens & Wooldridge, 2009).

Few studies have been conducted to compare among the numerous matching and weighting methods, and given the different estimands from each, the methods may not be comparable. Frolich (2004) in a comparison of pair, *k* nearest neighbor and inverse probability weighting, found inverse probability to be the worst performing among the three methods. Alternatively, Harder, Stuart and Anthony (2010) and Kang and Schafer (2007) suggested that the relative performance of matching or weighting methods is context dependent and sensitive to specification of the propensity score model. Because they estimate different effects—matching an ATT and weighting an ATE—they may not be comparable. In this study, inverse probability weighting was used to ensure that an ATE was estimated and to have higher efficiency and power.

Propensity Score Methods and Unconfoundedness

PSA has received substantial attention over the past decade in social and policy research. The approach remains controversial because of the fallibility of the unconfoundedness assumption in non-random settings, and much of the policy audience for typical evaluation studies may not find evidence from PSA research studies credible (Cook, Shadish & Wong, 2008). The efficacy of PSA in equalizing non-equivalent treatment groups, relative to experimental randomized designs, has generated substantial controversy (Cook & Steiner, 2009; Dehejia & Wahba, 1999; Langbeim, 2009; Pirog, Buffardi, Chrisinger, Singh & Briney, 2009), including specifically as the approaches relate to school evaluations (Cook,

1999; Schochet, 2005). For the results of any study using PSA to be interpreted as causal by the broadest audience possible, these controversies—meaning, the unconfoundedness assumption of the RCM—must be addressed to the extent possible within a specific study.

Smith and Todd (2005) demonstrated that PSA did not sufficiently address unmeasured confoundedness and although they may have potential in narrowly-defined situations the authors argue that PSA do not constitute a general application. Agodino and Dynarski (2004) and Wilde and Hollister (2007), using within-study replications of RCTs with PSA, demonstrated that PSA did not replicate the randomized study design. Other studies have demonstrated, alternatively, that PSA is a valid approach for inferring causality in observational studies as long as certain conditions and caveats are considered. Glazerman, Levy and Myers (2003) show that the coupling of PSA with regression covariate adjustment procedures—the first of which addresses covariances between treatment assignment and measured variables and the second of which addresses covariances between measured variables and the dependent variable—reduced bias more than either procedure alone (Glazerman, Levy & Myers, 2003). Diaz and Handa (2006) demonstrated that what they termed “off-the-shelf” variables that, in typical education data would include such demographic variables as race/ethnicity and free/reduced lunch, did not sufficiently capture the assignment process and that some effort must be taken to measure correlates of assignment. Cook, Shadish and Wong (2008), weighing the evidence from multiple studies, conclude that PSA is more likely to satisfy the unconfoundedness if pretreatment measures of the outcome can be included in the propensity score model. In sum, PSA may not constitute generalizable observational study methods and should be used sparingly when conditions are right.

PSA has often been applied in settings where the data are inadequate to model the treatment assignment process, leaving it unclear whether the unconfoundedness assumption was supported. The typical approach to establishing that unconfoundedness condition is satisfied in is to first demonstrate that observed covariates are balanced across propensity-score adjusted treatment conditions, and then to subsequently assume that this balance constitutes evidence that treatment groups are balanced on unmeasured covariates or that unmeasured covariates of the outcome do not exist. The first assumption implies that the observed covariates proxy for the unobserved covariates and no residual covariation remains (Rosenbaum, 2002). Certain researchers (e.g., Wilde & Hollister, 2007) have found neither of these arguments sufficiently credible.

My objective is not to take a stand on whether the balance argument is credible or not, but to offer a more rigorous criterion based on a weaker assumption. Specifically, PSA studies conducted in settings where a randomized control trial has also taken place offer the opportunity to replicate the RCT estimate on the same group of treatments but using a non-randomized PSA comparison group (Peikes, Moreno & Orzol, 2008). For the PSA estimand to be a valid causal effect in this scenario, I need only assume that the treatment intensity in the RCT treatment schools is the same during the non-random study period as it was during the RCT. If such an assumption holds, then evidence of the credibility of the PSA design will be conclusive. I argue that the present study may have such a design, for which I have limited quantitative and qualitative evidence to support the necessary assumption. This design therefore constitutes what Smith and Todd (2005) labeled a “mapping from the characteristics of the data and institutions available in particular evaluation contexts to the optimal non-experimental estimators for those contexts” (p. 397); in other words, although

PSA may not be applicable in all settings, it may be ideal in the present context. In this section, I present extant research on PSA as a replication method that demonstrates the weakness of relying only on balance in observed covariates and supports the strength of the within-study comparison design. These studies have some similarities to the present study design.

PSA as a Replication Method

Agodino and Dynarski (2004) compared experimental and PSA estimates of 16 dropout prevention programs on several student outcomes, including dropout, education aspirations, self-esteem, and attendance. An experiment showed that two of the programs were effective at reducing dropout. Using a difference-in-difference estimator to be consistent with the experimental study design, the authors show that the significant impacts observed in the experiment were not replicated. Non-significant findings in the experiment were also not replicated, with some ineffective dropout programs showing that they increased the odds of dropout. Using data from the Tennessee STAR class size randomized experiment, Wilde and Hollister (2007) similarly demonstrate that PSA do not give the same answers as the randomized experiment. Wilde and Hollister found that the PSA matched comparison groups were more equivalent to each other than the randomly assigned groups were on observable characteristics, but that unobservable characteristics had a heavy influence on the results in the PSA design.

Other researchers have reached similar conclusions. Peikes, Moreno and Orzol (2008), conducted a replication of an experimental study of an education and employment program. Using the results of a PSA model alongside the results of three randomized experiments using difference-in-difference and regression estimates, and using a statistical

test for the equality of the PSA and randomized designs, the authors compared the performance of these programs in reported earnings and proportion employed. To give an example of one of the three programs evaluated, the evaluation of the New York treatment showed that impact estimates for employment were similar across the two designs, but the estimated impacts on earnings were very different, with the experiment suggesting negative effects and the PSA suggesting positive effects. Peikes, Moreno and Orzol (2008) argued that PSA requires knowledge of an experimental impact estimate as the only criterion for the unbiasedness of a PSA estimate, which is typically a drawback as such estimates are not usually available.

Shadish, Clark and Steiner (2008), who took a different approach and designed a within-study comparison, argued that most of these failures of PSA to replicate RCTs were matters of design, not of PSA in general. The authors randomly assigning college students to an RCT or an observational study. Students in the observational study were permitted to self-select into their treatment. The treatment and comparison groups demonstrated non-equivalence before treatment and, further, without a PSA design the effects of the program being evaluated were different between the experimental and non-experimental designs. Using a series of covariate specifications, Shadish, Clark and Steiner (2008) show that typical demographic measures did not lead to bias reduction in PSA designs, but that additional measures less commonly available in administrative data but possibly collected via survey (e.g., indicators that measure motivation to participate) may help. In the case of their study, they were able to measure the variable determining self-selection into treatment. The findings demonstrated that if covariates are rich enough to measure the selection process or

proxy for it, PSA can replicate experimental studies. These findings were replicated by Pohl, Steiner, Eisermann, Soellner and Cook (2009).

Cook, Shadish and Wong (2008) conducted the most robust assessment of the quality of intervention program replication studies using PSA designs. Eight of the studies that they cited relied on PSA, including three education studies cited above, Agodino and Dynarski (2004), Wilde and Hollister (2007), and Shadish, Clark, and Steiner (2008). Cook, Shadish and Wong (2008) argue that collectively, the studies that they reviewed suggest that localized comparison groups should be used, and that richer data than typical convenient measures such as race/ethnicity and gender are required. Following Cook, Shadish and Wong (2008), other researchers have demonstrated that flaws in PSA designs are not inherent flaws in PSA but are context-specific. For example, Bifulco (2012) demonstrated that the extent to which pretreatment outcome measures are available and used, and the geographic homogeneity of the comparison pool for the matching or weighting, largely determine whether a PSA design was valid. The findings in these studies suggest the importance of a credible assumption of unconfoundedness and lend support to the idea of using the within-study replication approach to assess the credibility of this assumption in settings where such a design is possible, such as in the current study.

Study Objectives and Research Questions

The objectives of this study were to re-evaluate CareerStart using a non-random sample to answer questions about the scalability of the program and address concerns about flaws in the RCT. Given the non-random assignment to treatment and the voluntary nature of adoption of CareerStart in some of the schools, the treatment effect may be biased in favor of a larger and significant effect of the treatment if the PSA did not credibly establish

unconfoundedness. Thus, the corollary objective of the study was to use methods that support the credibility of the unconfoundedness assumption of the RCM in a non-randomized setting. The following four research questions are answered by this investigation. *First*, do students in schools receiving CareerStart perform better on end-of-grade (EOG) reading and math exams, as measured across all three middle grade levels, in comparison to students in comparable schools not receiving the treatment? *Second*, are the effect sizes observed in this non-experimental evaluation similar to the effect sizes observed in the RCT conducted between 2006-07 and 2008-09? *Third*, are there differences in effectiveness between schools in the home district and schools in other districts? *Fourth*, is the claim that the PSA design provides an unbiased and consistent estimate of the treatment effect valid?

Methods

PSA Design

I conducted a replication of the RCT design on the group of 20 schools in a non-random implementation of CareerStart, including 13 schools outside of the RCT district and 7 schools within (see Table 3.1A and 3.1B). As a propensity score design this study is unusually robust, because it also allows me to incorporate the seven original RCT treatment schools, allowing for a direct assessment of the credibility of the claim that assignment was conditionally unconfounded under a weaker assumption than the balance assumption typically requires. The design subsequently includes, and builds on, the interrupted time series model used in the RCT (Woolley et al., in press), adding to it a richer set of covariates at the student, classroom, teacher, and school level.

Several factors were taken into account when developing the propensity score model. First, the treatment was assigned to entire schools. In these settings, matching of treatment

schools to comparison schools would have been desirable, because the PSA model was intended to approximate the assignment process (Stuart, 2007). In this case, this would have meant matching or weighting the 20 non-random treatment middle schools (plus the seven RCT treatment schools) to the 350 other North Carolina middle schools, or a subset of these middle schools. In such a design, the PSA model would have consisted of pre-treatment characteristics of the middle schools, which would have included aggregates of student characteristics during the pre-treatment period. Information about the students enrolled in these middle schools during the study period (which I label the "study cohort") would not have been considered.

However, the analysis design called for modeling student performance over both elementary and middle school years. For a middle school PSA design to satisfy the requirement of equalizing the treatment groups, the middle school information used in the PSA model would have had to equalize over both the middle schools themselves and the elementary schools that fed these middle schools. Because the elementary school years for these students comprise the pretreatment period in this study, leaving their characteristics out of the PSA model may not have yielded equivalent groups of pre-treatment characteristics. An exploratory analysis conducted on several matched middle school samples demonstrated that this was the case, and that the two treatment conditions were not balanced. The propensity for a middle school to have been assigned to treatment did not indicate the propensity of any of the students in its constituent elementary schools to be assigned to a treatment, and due to migration of students many of the elementary schools fed into both treatment and non-treatment schools.

Instead, I used a second-best strategy of estimating a PSA model comparing students in the treatment schools to other students in the same cohort in other middle schools in the state. All of the students in sixth grade at the start of the 2008-09 school year were identified as an evaluation cohort, with the students in the 20 non-random treatment schools and seven random treatment schools identified separately from the students in 352 non-treatment middle schools. In total there were 84,968 students. This allowed for a larger and richer set of pre-treatment covariates for predicting assignment to treatment, including middle school pre-treatment covariates, elementary school covariates and student characteristics that could not be included a model strictly related to middle school assignment (see Table 3.3 for all covariates). This second-best design may not approximate the assignment mechanism, which in this case would mean modeling school assignment to treatment rather than student (Cook, Shadish & Wong, 2008). Rather, it approximates a design in which the treatment was assigned to schools and students subsequently assigned to these schools. Nevertheless, the results show that the design demonstrated improved balance when inverse probability weighting was used, relative to the balance across treatment conditions prior to weighting. Inverse probability weighting was selected to estimate an ATE for CareerStart, rather than an ATT that would be obtained via matching, as well as to minimize biases from loss of participants outside of the area of common support.

Data and Sample

Data. Three sources of data were used in this study. First, an existing database of North Carolina Department of Public Instruction (DPI) data, maintained by the Carolina Institute for Public Policy (CIPP), contains information on all schools and students in North Carolina dating back to 2004-05, four years prior to implementation in the replication sample

(2008-09). These data therefore include academic performance on sixth grade students in the entry cohort in 2008-09 dating back to third grade, the first grade in which end-of-grade testing was administered. These data also include school level data that may explain a substantial amount of the heterogeneity between treatment conditions in the years prior to schools joining the treatment condition.

Second, implementation data were provided by two sources. Quantitative data contained the number of lessons used by teachers in each subject in each school during the RCT and PSA evaluation periods. Qualitative data were collected in the form of an interview with the CareerStart Coordinator responsible for implementation of CareerStart in the two evaluation periods. The interview contained questions related to implementation intensity during the two evaluation periods in the multiple treatment subgroups.

Sample. The school sample consisted of three groups of treatments and one group of comparisons. The three treatment groups were labeled a) *former RCT controls*, consisting of the seven schools that initially served as controls in the RCT but were later admitted to treatment after a two-year wait, which can be considered a non-random assignment; b) *non-RCT treatment*, consisting of the group of schools to which CareerStart was disseminated non-randomly; and c) *RCT treatment*, consisting of the seven schools used to evaluate the efficacy of CareerStart in the 2004-2009 RCT, included here only as a check on the credibility of the design. There were middle schools in the RCT district that implemented CareerStart during one or more years but were excluded because the schools newly opened late during the PSA evaluation period or had been opened during the RCT but were closed during the PSA period. The funding of the position of the Coordinator was shared by the districts during the PSA period.

The schools in RCT control and non-RCT treatment can also be organized according to the staggered implementation timetable. The schools joining the treatment started using CareerStart according to the timetable shown in table 3.1. Table 3.1A shows the years in which the treatment could be evaluated in these schools, and Table 3.1B shows the corresponding grade levels for the cohort. Partitioning the treatments into four groups of schools addressed the staggering of the non-random implementation of CareerStart over the years 2008-09 (Group 1; 13 schools; sixth grade and later), 2009-10 (Group 2; 2 schools; sixth grade and later); and 2010-11 (Groups 3 and 4; 3 schools with a sixth grade starting cohort and 2 schools with a seventh grade starting cohort, respectively). Comparison students would be drawn from sixth grade students in 2008-09. Steps were taken to ensure that students and schools only appeared in one group. Table 3.1 also shows that data are currently available up to the school year 2010-11, which implies shortened study periods for groups 2, 3 and 4. Table 3.4 contains the unweighted description of the student sample.

Measures

The measures are organized according to their use in the analysis models, and in the descriptions I note the role that each variable played in the propensity score models.

Dependent variables. The dependent variables were end of grade (EOG) math and reading test scores, collected over all six available grade levels (third through eighth). This provided three pre-treatment time points and a time point for as many as three grade levels of middle school. North Carolina EOG exams are developmentally scaled measures, appropriate for assessing change or growth over time. Because the scales were revised during the study period, the outcomes were standardized. The addition of multiple pre-treatment observations in the third through fifth grade waves provided more reliable evidence of learning prior to

treatment, which strengthened the counterfactual by minimizing random fluctuation and regression to the mean (Raudenbush, 2001).

Time. An underlying "root" time variable ranged from -2 (corresponding to third grade) up to 3 (corresponding to eighth grade) with a center (zero) on fifth grade. This root time effect, which was entered as a predictor, served multiple purposes. First, as learning could be viewed as a developmental characteristic, this variable captured children's underlying learning trajectory (e.g., Willett, Singer & Martin, 1998). Second, this variable controlled for all unmeasured time-varying correlates of assignment and the outcome that were common to participants in both treatment conditions. A quadratic version of the root time variable was included to account for non-linear development (Willett, Singer & Martin, 1998).

Second, an increment to learning or an interruption in the trajectory occurred at the point that students transitioned to middle school. This interruption variable was a time-varying effect coded 0 for third through fifth grades and coded 1 for sixth through eighth. This interruption was a discontinuous jump in the trajectories of students between their fifth and sixth grade end-of-grade assessments. Reflecting the intense changes these students undergo during this transition, it was usually a negative value; that is, the average student's performance tended to go down when he starts middle school in sixth grade (e.g., Cook, MacCoun, Muschkin & Vigdor, 2008).

Treatment effects. There were multiple treatment effects tested in these analyses, reflecting the need to estimate an overall impact estimate across all non-randomly assigned schools (to answer questions 1 and 2); the need to estimate district-specific effect for

answering question 3; and to estimate a non-random treatment effect for the original RCT treatment group for answering question 4.

First, a main binary treatment variable captured the residual difference between students in treatment (=1) and comparison schools (=0) immediately *prior* to treatment. Due to the PSA design and interrupted time series, I hypothesized that this difference would be non-significant. This binary treatment variable was also the dependent variable of the propensity score model. Second, this binary treatment variable was interacted with the interruption to produce a difference-in-difference (DD) estimate, a difference between students in each condition in the difference between average MS and ES performance adjusted for development. The DD was interpreted as the focal effect for answering questions 1 and 2 regarding the magnitude and significance of the non-random treatment. It was hypothesized to be positive and significant to answer question 1, and to be statistically equivalent to the RCT treatment impact for question 2.

Third, the effects of the treatment in the RCT control and non-RCT treatment schools were estimated in two ways to answer questions 3 and 4. First, a variable indicating that the school was in the RCT district was interacted with the DD (producing a difference-in-difference-in-difference; DDD) treatment effect. The coefficient for this variable was the added benefit of the treatment experienced by students in RCT control schools relative to the treatment experience of those in non-RCT treatment schools (which would be negative if the non-RCT treatment schools had greater treatment effects), partly answering question 3. A second set of treatment variables enabled the estimation and testing of effects for each treatment group relative to the comparison group. This second set of treatment variables replaced rather than augmented the main treatment variable described above, and each was

interacted with the interruption to produce an independent DD treatment effect for each condition: former RCT control and non-RCT treatment for partly answering question 3; and RCT treatment for answering question 4.

Student, classroom, teacher and school covariates. Covariates were required for the propensity score model to predict the log-odds of treatment. Further, as Glazerman, Levy and Myers (2003) demonstrated, PSA designs can be made more credible in addressing the unconfoundedness assumption by including covariates in a regression model for analysis of the study outcomes. The CIPP database contains numerous variables including student demographics, test score data over multiple years, and indicators of learning status such as limited English proficiency; classroom variables that may capture the confounding influences of peer learning effects on the treatment effect; teacher variables that capture student-to-student differences in teachers' skill level and readiness to teach their content areas; and school variables that account for aggregate demographics and may proxy for effects related to geography and local economic conditions. Table 3.3 lists all of the variables that were used during the imputation, propensity score and analysis stages of this study, and indicates in which stage of the study each variable was used.

Data Analysis

Propensity score model. A logistic regression as described in model *1* used a vector of student, classroom, teacher, and school characteristics to predict the log-odds of treatment assignment. The chosen covariates in the PSA model should be predictive of both treatment assignment and the outcome and rich enough to proxy for unobserved covariates of treatment and the outcome, and they should be balance-promoting (Diaz & Handa, 2006; Imbens & Wooldridge, 2009). The outcomes themselves cannot be used for selection, but proxies can

be used instead (Kelcey, 2011). Fifth grade math and reading scores were used as proxies for the middle grades outcomes. For selecting covariates and nonlinear functions of covariates that were highly predictive, numerous approaches have been recommended (e.g., Dehejia & Wahba, 1999; Guo & Fraser, 2009; Hirano & Imbens, 2001; Rosenbaum & Rubin, 1984). I synthesized an approach from these recommendations that fit the unique needs of this study: a high number of comparison participants; a high degree of imbalance between treatment and control samples (approximately 1:15), and a large vector of potential student, elementary school and middle school covariates (more than 250). First, I calculated the correlation between all available variables and the outcome proxies to narrow the number of variables for the propensity score model to 116 variables (these are labeled in table 3.3 as *candidates* for the PSA model).

Following this reduction, exploratory logistic regressions were run to eliminate variables that contributed little, conditional on the other variables, to the prediction of treatment assignment. Among the variables entered into the PSA model were student math and reading scores in fifth grade, as well as differences between students third and fourth grade math and reading scores and their corresponding fifth grade scores, and both elementary and middle school average student performance during the year in which these students were in fifth grade. These additions address the findings of studies such as Cook, Shadish and Wong (2008) that multiple pre-treatments may aid in eliminating the influence of unmeasured covariates of the outcome and the treatment. With a very large sample size (consisting of 84,968 students), I assumed that the model could contain many variables and that variables should be eliminated only if they were collinear or were at risk of causing complete or quasi-complete separation of data points in the logistic regression. Collinearity

was detected using the variance inflation factor (VIF) statistic in linear regression models. Variables with high VIFs (e.g., VIF = 20) were removed, and most variables with VIFs between 10 and 20 were also removed. Minimizing the collinear relationships helped to reduce the occurrence of complete separation (linear dependence) or quasi-complete separation, which were tested via trial and error. Complete separation was reported by SAS Proc Logistic as an estimation failure. Quasi-complete separation was indicated by unusually high log-odds coefficients. Because of the time involved in the trial-and-error estimation of the PSA model, these exploratory analyses and balance tests were estimated on non-imputed data.

Third, these variables were subjected to a search for potent interactions and quadratics, and a small group of interactions and quadratics that had satisfactory Wald χ^2 statistics were retained for testing. The search considered all possible combinations of covariates, including quadratics, and was evaluated using a modest Wald criterion (χ^2 = 1.4). At all stages, models showing evidence of complete separation or quasi-complete separation were either reduced by removing variables with high coefficients and errors (prior to the search), or the stage was discarded altogether (during the search).

After the propensity score regression was run, a method for adjusting the sample using the propensity scores was needed. Because the objective was to calculate an ATE for CareerStart similar to that of the RCT, I used inverse probability weighting. Weights were calculated as $g(1/P)$ for treatment school students and $g(1/(1-P))$ for comparison school students; $g(.)$ was a re-norming function such that the post-weighting sample size was equal to the pre-weighted sample size (Hill, Weiss & Zhai, 2011). The weights were trimmed at 20 (Imbens & Wooldridge, 2009). The propensity score model was a pooled binary outcome

model treating all student, elementary school and middle school characteristics as independent.

Balance on covariates between the two treatment conditions was subsequently assessed. Two balances were estimated, one without weighting to determine the pre-adjusted differences between groups, and the second with the weights from the PSA model, and a comparison between these findings was conducted to ensure the PSA model improved, rather than worsened, balance (Harder, Stuart & Anthony, 2010). Imbens and Wooldridge (2009) recommend a clinical difference be used to assess whether groups were balanced after regression, as the clinical difference is insulated against the influence that sample size has on statistical significance. The clinical difference was calculated as:

$$(2) \qquad T = \frac{(\bar{X}_T + \bar{X}_C)}{\sqrt{\left(\frac{S_T^2}{N_T}\right) + \left(\frac{S_C^2}{N_C}\right)}}$$

$\bar{X}_T$ and $\bar{X}_C$ were the averages in the treatment and comparison groups; S_T^2 and S_C^2 the variances in each group; and N_T and N_C the number of students. To be as robust as possible, I assessed balance on two groups of variables—those used in the PSA model and those that were found to be candidates but removed from the PSA model.

Analysis model. I used a multilevel or hierarchical linear model (HLM; Raudenbush & Bryk, 2002) to estimate the interrupted time series model. HLM was used because treatment was implemented at the school level, although the outcomes were student level measures observed over repeated time points. Consequently, in this model there were three levels of data: Schools (J schools sampled); students (I_j students in each school j); and time (multiple measures w_t on each student i_j). According to this design, the treatment variable was the product of a school level assignment interacting with a time-level variable.

Modeling change over time on individuals (students) nested within clusters (schools) implied a complex error variance structure due to the simultaneous influence of dependence between multiple observations on each individual and between students within each school (both of which manifest as non-zero off-diagonals in the error covariance matrix). Not accounting for these sources of dependence would lead to under-estimated standard errors and spurious findings of significance. In the context of the PSA model employed in this study, in which comparison students could be sampled from any of the 350 non-participating middle schools in the state, this concern was not as great as it was in the RCT, but nevertheless the multilevel design of the Woolley et al. (in press) study was left intact. For questions 1 and 2 the model was estimated as follows:

$$(3)\ Y_{isw} = X_{is}\beta_i + X_s\beta_s + \beta_{MS}MS_w + \beta_{CS}CS_s + \beta_{DD}CS * MS + u_{i0} + u_{s0} + u_{i1} + u_{s1} + e_{isw}$$

Assume Y_{isw} is math or reading achievement in each grade level; $X_{is}$was a vector of student characteristics; X_s a vector of school characteristics. MS_w was the transition indicator; CS_s the CareerStart treatment assignment variable in the PSA design; and $CS * MS$ the treatment difference-in-difference interaction. The u_{i0}, u_{s0}, u_{i1}, u_{s1}, and e_{isw} terms were errors at the student (intercept and MS slope), school (intercept and MS slope) and time (residual) levels. The RCT treatment group was left out of this model. For question 3D the following model was estimated:

$$(4)\ Y_{isw} = X_{is}\beta_i + X_s\beta_s + \beta_{MS}MS_w + \beta_{CS}CS_s + \beta_{DD}CS * MS + \beta_{DDD}CS * MS * D + u_{i0} + u_{s0} + u_{i1} + u_{s1} + e_{isw}$$

D = 1 indicated a school was in the RCT district, and $CSxMSxD$ was the difference-in-difference-in-difference for estimating the difference between RCT and non-RCT effects.

The RCT treatment group was also left out of this model. For question 3S the following model was estimated:

$$(5)\ Y_{isw} = X_{is}\beta_i + X_s\beta_s + \beta_{MS}MS_w + \beta_{RCTC}RCTC_s$$

$$+ \beta_{NRCT}NRCT + \beta_{RCTT}RCTT + \beta_{DDRCTC}RCTC_s * MS + \beta_{DDNRCT}NRCT_s * MS$$

$$+ \beta_{DDRCTT}RCTT_s * MS + + u_{i0} + u_{s0} + u_{i1} + u_{s1} + e_{isw}$$

$RCTC$ = 1 indicated a former RCT control school; $NRCT$ = 1 indicated a non-RCT treatment school; and $RCTT$ = 1 indicated an RCT treatment school. All comparison schools had a value of zero on all three of these parameters. Each of these was interacted with the transition variable MS to produce difference-in-difference treatment effects for each treatment sub-group as shown in the model.

Comparison with RCT estimate. Two of the research questions (2 and 4) implied a statistical test between the non-random estimates from this study and the estimate from the RCT. To conduct the test, a standard error had to be estimated. Although Bifulco, (2012) recommended a standard error estimated from a function of the RCT and PSA variances, the PSA variance was found to be too large owing to the large sample size in this study, which rendered all hypothesis tests non-significant. Consequently, I used the standard error from the RCT, which was 0.039 for math and 0.018 for reading.

Estimation procedures. Generalized least squares is an estimation approach that accommodates failures of independence and enables decomposition of variance according to approximate source of time, individual or school. Proc Mixed in SAS 9.1, estimated generalized least squares models with restricted maximum likelihood. Because of the sample size and imputation, maximum likelihood was computationally intensive. A linear approximation method, *mivque0*, which provides similar estimates in large sample sizes, was

used. To ensure that the range of values modeled was realistic for all of the participants and their schools, that each variable for each school could realistically have a value of zero (such that the random intercepts and slopes could be interpreted validly), and that the coefficients of the models could be interpreted as the effect for the average student, all variables (Table 3.3) were mean-centered.

Implementation data analysis. Quantitative implementation data on CareerStart lesson usage was reported at the school-by-subject-by year level, for each evaluation period. RCT treatment school implementation data were available from 2006-07 to 2010-11, and RCT control school implementation data in 2009-10 and 2010-11. For the PSA evaluation period, only the first two years of data were available. These data were subjected to a t-test comparing the average count of lesson use in each evaluation period. The quantitative data was supplemented by an interview conducted with the CareerStart Coordinator in February 2013. Notes taken during the interview were reviewed for statements by the Coordinator regarding the relative intensity of implementation in each evaluation period and treatment subgroup, as well as statements regarding demands on the Coordinator's time and differences observed by the Coordinator in implementation in each school. Intensity was expressed by the Coordinator primarily as its inverse in the form of the amount of demands on her time during each evaluation period.

Missing data and imputation. There were missing data at the student level on a large number of characteristics, and missing school level data for years in which the schools were not open. Missing values were tested for completely random missingness using the Little (1988) chi-square test, a multivariate test of whether students with missing values were

different from students without missing values. This test was rejected, indicating that at least some of the data points were conditionally missing.

An imputation model containing all of the PSA model and analysis variables was run. Five imputations of the data were simulated using SAS Proc MI. All analyses were conducted on each imputation separately using SAS and then the parameter estimates from these distinct analyses combined according to rules developed by Rubin (1976). A version of the PSA and analysis models without imputation was also run to assess the robustness of the findings to the imputation.

Robustness diagnostics. I reconsidered several subjective decisions using a robustness diagnostic to assess their impact on the findings (Table 3.5 summarizes each robustness model relative to the main model). These included diagnostics related to missing data (discussed in the missing data section). For group 4 schools, the treatment was not administered in sixth grade. Thus, these schools were left out of the main models. In a robustness test, they were included and because they were not in treatment in sixth grade, the treatment period in these schools absorbed a non-treatment year, which I hypothesize would downwardly bias the treatment impact. The shorter post-treatment period in groups 2 and 3 presented another opportunity for treatment effect dilution, and I estimated robustness models with these groups removed. Finally, following Woolley et al. (in press), students who moved during the study period (17.5% of the sample) were excluded from the analysis, and the model retested.

Results

Propensity Score Model

Selection of covariates for PSA model. In the first stage of the PSA model, I selected covariates of the outcomes that were predictors of treatment assignment, using fifth grade math and reading achievement (prior to treatment) as proxies for the outcomes. Given the large sample size, I allowed for variables with correlation coefficients as small as 0.01 on either math or reading to be considered as candidates. The variables that satisfied this condition, and their correlations with math and reading achievement, are shown in Table 3.6. The second stage of the PSA design was to enter these covariates into regression models with treatment assignment as the dependent variable and test for collinearity and separation. Variables with high VIFs (e.g., VIF = 20), indicating collinearity, were removed. Variable combinations that produced complete or quasi-complete separation, were detected by trial and error using SAS Proc Logistic. A set of 65 variables, which included 16 student level characteristics, 34 elementary school characteristics, and 15 middle school characteristics, was selected using this approach. Using an iterative search procedure I described in the analysis section, quadratic and interactions terms based on these 65 variables were tested. However, the quadratics and interactions that emerged in testing did not subsequently improve the balance of the treatment conditions and were discarded from the final model.

After estimating the final PSA model (Table 3.7), box plots and kernel density functions were plotted across the whole sample (Figures 3.1 and 3.3) and by the treatment conditions (Figure 3.5). After trimming the propensity scores in the treatment and control such that the weights applied were no greater than 20, these plots were redone (Figures 3.2, 3.4 and 3.6, respectively). Generally, the plots show a tendency for students who were in

schools not in the treatment to have near-zero propensities, though near-zero propensities also occurred in the treatment condition, though less often by a factor of 150.

Balance

Covariate balance before and after inverse probability weighting were compared to demonstrate whether the PSA model reduced measured differences between treatment groups. Covariate balance before the PSA model, using the variables in the final PSA model, showed that 18 characteristics were significantly different (Table 3.8). All 18 of these characteristics had sufficiently improved balance after the PSA that they were no longer significantly different (Table 3.10). An additional six characteristics from the original list of candidate covariates that were not included in the PSA model were shown to be significantly different before the PSA (Table 3.9), with three of these improved such that no significant differences remained after the PSA (Table 3.11).

One characteristic in the PSA, proportion of Nationally Board Certified middle school teachers, went from being not significantly different (Table 3.9; d = -0.14) to significantly different (d = -0.28) after the PSA (Table 3.10). Three variables not in the PSA that were found different prior to the PSA (Table 3.9) remained significant after the PSA (Table 3.11), including per-pupil spending on district administration (d = 1.63 before PSA, and d = 0.43 after PSA) and school is located in a large city (d = -0.34 before and after PSA), and proportion of teachers returning (d = 0.33 before PSA to d = 0.26 after PSA). An additional variable, elementary school has year-round calendar was not significant before the PSA (d = -0.23) but significant afterwards (d = -0.27). The inverse probability weighted student sample characteristics are reported in Table 3.12. The PSA model covariate that remained

unbalanced after the PSA, proportion of National Board Certified teachers, was not a significant predictor of math or reading achievement in any models (Tables 3.14-3.19).

Analysis Model

The results are organized according to the four questions of this study, and robustness results are reported separately. Table 3.13 contains a summary of the treatment effect results, for main and robustness models, and Tables 3.14-3.19 contain the full results for the main models.

Questions 1 and 2. Question 1 concerned the overall impact of CareerStart, implemented non-randomly in 20 schools, on math and reading end-of-grade achievement using a PSA design. The RCT treatment schools and 2 non-RCT treatment schools that started implementation in seventh grade were excluded from this analysis. On math achievement (Table 3.13 and 3.14), the impact was negative but not significant ($b = -0.04$). For reading (Table 3.13 and 3.17), the impact was significant and positive ($b = 0.04$, $p < .05$).

Question 2 concerned whether the findings from this PSA study were statistically different from the findings in the CareerStart RCT. For math, the findings were significantly different (Table 3.13). The CareerStart RCT effect for math was 0.039 standard deviations, yielding a difference of -0.079 ($t = -7.9$; $p < .001$). The difference for reading was not statistically significant (Table 3.13). The RCT effect for reading was 0.018 standard deviations, yielding a difference of 0.022 ($t = .96$; $p = .339$).

The finding for reading—that the RCT and PSA effects were not statistically different, and that the reading effect in the PSA design was significant—may support the argument that the RCT was underpowered for detecting reading treatment effects. In fact, the reading standard error was the same (SE = .02) in both designs. However, the PSA impact

was more than twice as large, which implies the relative intensity may have been higher. This higher relative intensity could be due to a number of factors, such as contamination of the control during the RCT or to a poor PSA model. Alternatively, it could simply have been a higher intensity implementation. To understand this, the PSA impact was examined for sub-group differences and effects.

Question 3. Question 3 addressed two related issues. The first (Q3D) was whether there was a difference between schools in the RCT district (*Former RCT controls*) and schools outside this district (*non-RCT treatment*) in the PSA treatment effect. The second (Q3S) was whether either of these two treatment sub-groups had a significant impact on math and reading achievement relative to the comparison group. Both of these questions were important in their own right, as one aim of this study was to determine if there were implementation peculiarities in the RCT district that were not replicated elsewhere. However, looking at sub-group effects may also help to understand why the reading effect was more than twice as large in the PSA as in the RCT design as shown for Q2.

For Q3D, which was tested using a DDD, the differences between the treatment impact for the Former RCT controls and non-RCT treatment schools was not significant in either math (Table 3.13 and 3.15) or reading (Table 3.13 and 3.18). Despite not being significantly different, the effects could still be different, and have different precision. For Q3S, there were no significant effects for math in either treatment subgroup (Table 3.13 and 3.16). For reading (Table 3.13 and 3.19), the treatment in the RCT control subgroup ($b = 0.03$) did not have a significant impact on reading achievement but the treatment in the non-RCT treatment subgroup had a significant effect ($b = 0.05$; $p < .05$).

In sum, the findings of Q2 and Q3 suggest that intensity of the treatment, particularly in language arts, may have been greater in the PSA study period, and that this greater intensity may have been even greater in the non-RCT treatment schools. Quantitative data based on CareerStart lesson use in the former RCT control schools compared to the RCT treatment schools shows that lesson usage in all subjects in the former RCT control schools was higher, with a difference of four lessons on average ($t = 2.05$; $p < .05$). A similar test on language arts lessons showed the same difference of four lessons, but the effect was not significant, likely owing to the smaller sample size.

Lesson usage statistics were not available in the non-RCT treatment schools, but qualitative data suggested a more complex picture of implementation. The number of schools to which the Coordinator was responsible for implementation was much larger during the PSA period, and included all seven RCT treatment schools, all seven RCT control schools, and 18 of the 20 non-RCT treatment schools. The Coordinator reported that this put a strain on her ability to implement CareerStart during the PSA period with similar intensity to that in the RCT. However, the Coordinator also indicated that schools in two counties demonstrated a high level of enthusiasm for the program. When these two counties were examined for their own DD effects—separately estimating parameters for the schools in these counties—one of the counties had an effect of 0.10 for reading achievement ($p < .05$), and no other effects were significant.

Question 4. The last question, regarding the comparison between the PSA impact estimate and the RCT impact estimate for the RCT treatment schools, was devised to address the credibility of the PSA model. Based on the assumption that treatment intensity was equivalent in both designs, a non-significant difference on a t-test comparing the findings in

each design for these schools would represent evidence that the PSA model credibly satisfied the unconfoundedness assumption. For math (Table 3.13), this test was rejected ($t = 3.9$; $p < .001$), implying that the PSA model may not have sufficiently eliminated common confounders of treatment assignment and math achievement. Alternatively, for reading (Table 3.13), the test was not rejected, implying that the PSA model may have sufficiently controlled for common confounders of assignment and reading achievement. Although different findings for math and reading on the credibility test were not hypothesized, the two subjects respond differently to policy interventions (e.g., Harris & Sass, 2011; Carlson, Borman & Robins, 2011) and often have different predictors such as those shown in Table 3.6 (e.g., elementary school classifications as low or no growth) and thus different PSA models for addressing confoundedness, implied by these findings, may be appropriate.

The assumption supporting this test, that treatment intensity in the RCT treatment schools was the same in the PSA period as in the RCT period, was examined using both quantitative and qualitative data. Quantitative data consisting of the total number of CareerStart lessons used in each subject and school during each evaluation period (Table 3.20) were compared using a t test and the null hypothesis of no difference was not rejected ($t = 1.34$; $p = .18$), indicating that intensity was statistically equivalent in both periods. Qualitative data, collected in the interview with the CareerStart Coordinator, lent support to the quantitative finding. By the time the PSA evaluation started, RCT treatment teachers would have been in their fourth year of implementation. The Coordinator reported that although the number of schools to which she was responsible was greater, the RCT treatment schools were sustaining the treatment with less direct oversight.

Analysis Model Robustness

Several additional tests (summarized in table 3.13) were run to ensure the main analysis model was robust. For math, all of the robustness tests – the non-imputed results, the models with group 4, the models without group 3, the models without groups 2 and 3, and the model without movers (students that changed middle schools)—agreed with the main model. For reading, there were some differences across these robustness tests. On the NI data, the models with group 4 schools, without groups 2 and 3, and models without movers, the overall reading effect was not significant. However, the reading effect was robust to removing group 3 schools. The same was true of the significant non-RCT treatment effect, which went up slightly when the group 3 schools were removed. There were no difference across the variations in robustness for the comparisons of the PSA impact with the RCT impact, likely as a result of the PSA impact varying within a narrow range across these robustness tests (e.g., from 0.02 to 0.04). On the PSA credibility test, the robustness findings also produced no difference, though the model for the RCT treatment schools was identical across several of the variations.

Discussion

In this study I used a non-random dissemination of a teaching practice intervention to schools inside a county in which an RCT of this program was conducted as well as schools in neighboring counties to retest the efficacy of the treatment relative to a set of comparison schools weighted by the inverse of the propensity score of being assigned to treatment. This design also allowed me to test the relative magnitude of the effects inside the RCT county and outside of this county, as well as to test the credibility of the PSA model. The results were not a match to the results of the RCT. In math, the treatment effect was not significant

overall, which was an unexpected change from the RCT finding that CareerStart promoted higher math achievement. This finding was robust to all subgroups and to all robustness variations that were examined. However, the analysis for question 4 testing the credibility of the PSA design showed that for math, the PSA did not credibly eliminate all confounders of assignment and math achievement. As a consequence, nothing can be said about the effect of CareerStart in promoting math achievement in the PSA, and I downplay the non-significant finding in this discussion.

For reading, alternatively, the treatment effect was significant, and contrary to expectations, it was driven primarily by non-RCT treatment schools, for which a significant subgroup treatment effect was found, rather than by former RCT control schools. Further, informed by information provided by the CareerStart Coordinator, I conducted a further analysis that showed that the magnitude and significance of the non-RCT treatment effect may have been due to enthusiastic implementation in one district. Finally, the PSA impact was more than twice the RCT impact. A number of explanations are possible for these findings.

My initial expectation that this study would demonstrate that the RCT was underpowered for detecting reading achievement effects was not realized. In fact, the standard error for reading achievement was equal in both designs (SE = 0.02). Because this design contained 379 middle schools, an underpowered RCT would have easily been detected by this comparison, manifesting as a smaller standard error in the PSA. Instead, the standard errors across the two designs were nearly identical and the significant and positive finding for reading in the PSA was driven by the greater magnitude of the effect, which was twice as large in the PSA. Although this difference in magnitude was not enough for it to be

found as a significant difference between the PSA and RCT estimates of the reading effect (question 2), it nevertheless rules out that the RCT was underpowered.

To understand what was behind this larger overall effect, I considered question 3 to examine the subgroup effects. An added unexpected finding was that non-RCT treatment schools had a larger effect that the former RCT control schools. The non-RCT treatment effect was 0.01 standard deviations higher, which although not a significant difference was enough for the subgroup effect to emerge only for the non-RCT treatment schools.

There are a number of potential explanations for the larger effect size in the non-RCT treatment schools, and accordingly, for the higher overall PSA effect. The first explanation, that the PSA was poor and the comparison group not equivalent, can be ruled out by the credibility test, which showed that equality of the RCT and PSA findings had a low probability of occurring by chance. A second potential explanation was that the RCT controls were contaminated during the RCT because of their indirect exposure to the treatment, which depressed the observed effect during the efficacy trial. Measures of fidelity such as lesson use statistics were not collected from the control schools, but indirect evidence exists that controls may have implemented CareerStart lessons. CRI use during the RCT was not significantly different in language arts, science of social studies (Rose et al., 2012), suggesting that some level of contamination may have occurred in these subjects and thus this explanation cannot be ruled out.

Another explanation is that, in fact, the treatment was more intense during the PSA evaluation period. Quantitative data comparing lesson usage in the former RCT control schools in 2008-09 and 2009-10 with lesson usage in the RCT treatment schools from 2006-07 to 2008-09 indicated that on average lesson usage was significantly higher (by 4 lessons)

in the former RCT control schools. Such information about the non-RCT treatments was not available. However, this does lend some support to the argument that treatment was more intense during the PSA period. Qualitative data present a mixed picture of implementation intensity. An interview with the CareerStart Coordinator revealed that she was responsible for four times as many schools during the PSA period, which put a strain on the Coordinator to implement CareerStart during the PSA period with similar intensity to that in the RCT. This would not have supported higher intensity implementation. However, in districts demonstrating a high level of enthusiasm for CareerStart, as noted by the Coordinator, implementation may have been more intensive, leading to a larger effect in some of these schools.

However, using lesson counts and the Coordinator's time as measures of fidelity may only narrowly depict the differences in implementation in the two periods and subgroups. The Coordinator's effort, although an important part of CareerStart, was not the only component to the intervention, and it was possible that the other components were implemented more intensively. For example, although the Coordinator's time for each school was less, it was possible that she may not have needed as much time per school after having implemented CareerStart for several years. That is, the Coordinator may have experienced learning effects that would have improved her effectiveness in implementation. Unlike lesson usage, no measures were collected on these aspects of treatment. Principal and teacher buy-in was the other component that was of concern in this study, largely to understand the role that the superintendent in the RCT district had in promoting the program. As reading outcomes were greater in the non-RCT treatment schools, it cannot be argued that the superintendent's support biased the RCT treatment effect upward, and may have biased it downward; it may

also indicate that buy-in was successfully obtained. Qualitative data support this argument. The Coordinator indicated that strong efforts were made to get principals and teachers in the new districts to buy in to the program, including having the RCT district superintendent involved in selling the program.

Limitations

There were several limitations in this study. First, the study used a non-randomized design, and the PSA could not be implemented to reflect school assignment to treatment, which I documented extensively in the methods section and implemented a second-best approach. Further, although the PSA was shown to be credible for reading achievement, it was not credible for math achievement. It may be possible that interactions of covariates of achievement in math were overlooked, and it may also be possible that so-called "off-the-shelf" variables such as those in administrative data files like the one used in this study, are not capable of adjusting for unmeasured confounders of math achievement. It was also possible that the proxy for middle school math achievement, fifth grade math, may not have been adequate in representing variation in middle school math achievement, as math scores tend to fluctuate more from year to year. In contrast, reading scores tend to be more stable from year to year. In a follow-up analysis, I re-analyzed the propensity score model using average elementary school math over third to fifth grade and student change in math achievement as measured by a linear growth parameter. There were no substantive changes to the balance between treatment conditions or the question four credibility findings. A different approach to incorporating interactions in the PSA models was also tried, but resulted in no substantive changes to balance or the credibility of the PSA. This suggests that

the variables typically available in an administrative education data set may not be adequate for controlling for unobserved covariates of math achievement.

A second limitation was the voluntary nature of program adoption in the 13 schools outside the RCT district, which may raise questions about the generalizability of study findings (Heckman & Smith, 1995). Because districts elected to receive CareerStart services (i.e., they effectively volunteered to be members of a study in which they had 100% chance of being assigned to treatment), selection effects could be present. The extent of this bias is not known.

Third, several schools in the treatment group were admitted late to treatment and therefore had limited post-treatment data. Two additional years of data collection would be needed to fill in seventh and eighth grade achievement for group 3 schools, and one additional year for eighth grade achievement for group 2 schools. Group 4 schools, in addition, were not treated in sixth grade, which was inconsistent with the treatment as designed. A further limitation had to do with the size of the sample (84,968 students). As a result of this size, several approximation procedures were used. These included using non-imputed data to conducted exploratory analysis on the PSA model, and using a linear approximation to maximum likelihood to estimate the final outcome models.

Fourth, there were missing data at both the student and school levels. Multiple imputation was used to impute the data five times. Although a larger number of imputation would be better, five was chosen due to the overall size of the sample being very large and computationally intensive. Fifth, the ability to understand the replicability of CareerStart was challenged by not having data on certain components of the implementation, such as the frequency and attendance at meetings, or the information shared across schools through the

Coordinator. To a certain extent, some of these measures represent factors important to the treatment that were not understood at the outset of the efficacy study, and some of them (e.g., the sharing of information across schools) would have been challenging to collect. Most of the components of the program are implemented to a certain extent through the responsibilities of the Coordinator, and thus would have reflected the Coordinator's reduced role in the expanded group of treatment schools, but nothing certain can be said. Although the limited implementation information painted an incomplete picture of the challenges associated with replicability, and information about components central to the success of the program will be important to obtain, the finding that the effect had greater magnitude in the non-RCT treatment schools is beneficial, suggesting that an effectiveness evaluation is appropriate. Finally, EOG test scores are widely understood to be imperfect measures of student learning, and caution is warranted when interpreting findings from test score data.

Implications for Policy

This study cautiously suggests implications for policy, including CareerStart as a model for encouraging adoption of relevance as a teaching strategy, as a program that may be scalable pending modest changes to implementation, and as a program that should include language arts instruction. First, this study complements other studies of CareerStart, conducted in an efficacy RCT, showing that CareerStart has significant and sometimes lasting benefits on student achievement (Woolley et al., in press; Woolley et al., revise & resubmit), working through motivational processes (Rose, unpublished manuscript) and student engagement (Orthner et al., 2012). Importantly, the findings that address questions 1 and 2 of this study indicate that the program continued to have a significant impact, though the findings indicated a different content area (reading achievement) than in the RCT (math

achievement). As noted in the limitations, this study can only partly address why the program continued to have an effect, and why the subject on which this effect was estimated was different; further, as I note in the research implications section that follows, the subject-specific effects and the role of each subject in CRI need to be understood better.

Nevertheless, CareerStart has promise as a model program for encouraging teachers to use career examples among other relevance approaches in their lessons. For education policymakers hoping to improve student achievement, CareerStart, and more generally, motivational practices like CRI, may be a low-cost method that complements accountability efforts. This argument is supported by evidence from this study that local conditions did not play a significant role in the RCT CareerStart effect. Contrary to my expectations, the personal investment of the RCT superintendent in the success of CareerStart, if such personal investment existed or had any effect at all in the RCT, did not upwardly bias the treatment effect in the RCT design relative to the non-RCT treatment schools, as the non-RCT schools showed a larger effect. The unique local conditions of implementation (Elias, Zins, Graczyk & Weissberg, 2003; O'Donnell, 2008) were an important concern that the program—through buy-in of principals and teachers—and this evaluation were intended to address. As a consequence of these findings, despite having incomplete implementation data in other areas, the local conditions may not have played as much of a role as I hypothesized. Through buy-in, the replicability of the program was promoted. This provides some initial evidence – though inconclusive until an effectiveness trial is implemented—that CareerStart is a model program for systemic change in instruction, an important goal of education research.

Other findings help make the case that CareerStart may be scalable to more heterogeneous populations of schools, teachers, and students, provided certain conditions are

met including making some changes to CareerStart. One of these conditions is to fund a position for a CareerStart Coordinator who is responsible for implementation of the program. The responsibilities of the Coordinator include helping principals and teachers buy-in to the value of the program; holding meetings with principals and teacher at multiple times during the year, encouraging use of the lessons through coordinated lesson planning sessions; encouraging the development of new lessons through the distribution of a newsletter describing local labor market careers; and acting informally to transmit information about the program between schools. It is not clear how many schools a Coordinator could effectively serve. In this evaluation, however, a single full time Coordinator was able to serve most of 27 schools spread out over 8 districts, though not necessarily as effectively as when only serving seven schools. By sharing responsibility for a single Coordinator, multiple districts could pool resources to fund such a position, such as the strategy used in the non-random replication that was the focus of this study, but fewer schools may enable the Coordinator to be more effective. If funding a new position becomes a problem, the Coordinator's responsibilities could be integrated into an existing district position. The Coordinator's responsibilities could be aided by encouraging teachers in schools implementing CareerStart with high fidelity to "sell" the program to colleagues. Because the effects outside of the RCT district were driven largely by the enthusiastic implementation in one district, this suggests further the importance in future scaled-up implementations of calling on teachers in such a district to help train other teachers and disseminate the program more widely.

While some changes were suggested, this study also suggested that some components may not need to be altered. First, the question about whether CareerStart should continue to include language arts lessons has been answered in the affirmative. In this evaluation,

CareerStart increased reading achievement by 0.04 standard deviations. In this population, the average growth during a middle school year was 0.02 standard deviations, which means that the treatment was more than double the typical annual rate of increase. More generally, it is possible that the cumulative effect of career relevant lessons across the four subject areas promoted reading achievement in this study or math achievement in the RCT. At present, this is not known. Second, the design of the packaged lessons by teachers in the RCT district may not have been central to the success of CareerStart in the RCT, as the findings of this study showed greater effects outside the RCT district.

Implications for Research

The research implications of this study pertain to the continuing effort to evaluate CareerStart, but also extend more generally to the use of PSA to conduct evaluations of educational interventions including teaching practice interventions.

Implications for CareerStart evaluation. The findings of this study, which showed that CareerStart had a higher intensity implementation outside of the RCT district, suggest a direction for future evaluations, including collecting data on each of the seven components of the treatment. Although program and evaluation documentation data support high fidelity in several areas (presence of coordinator, staggered implementation, training, and planning meetings), teacher and curriculum coordinator buy-in, and teacher supplementation of the lessons have not been studied. Further, evidence for a social network with the Coordinator at the center, sharing information across schools, was available only via interview data.

I also argue that the experiences of the CareerStart Coordinator should also be studied. One suggestion I have made regarding the higher implementation in the PSA evaluation period, despite the larger number of schools, is that the Coordinator experienced a

learning effect and became better at her job over the five years of these combined studies. The evidence in this study do not clearly indicate that Coordinator learning effects exist, and the program has no theory of Coordinator learning effects built into it, but I speculate that such learning effects are fairly intuitive and may explain why intensity remained high despite having four times as many schools. If learning effects existed, then when the program entered its fourth year of implementation, the Coordinator should have been much more effective in her role. To put it another way, the findings suggest that it may take some time for a new Coordinator to learn the best approaches to program implementation. Just as important, the feasibility of incorporating the Coordinator's responsibilities into an already-existing district position should be examined. This would be a cost-effective approach, but may result in poor implementation as the Coordinator's duties would conflict with duties of the existing position. The present study provides some evidence, alternatively, that multiple districts could pool resources to share a Coordinator. In a future evaluation, multiple versions of the treatment could be tested in this way.

Finally, the findings of Tatar and colleagues (2008) suggest that the packaged design of CareerStart may support its replicability in other settings, but the effect on teacher buy-in of having the pre-packaged lessons developed by teachers in the RCT district has not been studied. One question that appears to be partly answered is whether to consider lesson design as a part of the implementation. Teachers in the RCT district designed the lessons, but no such design was incorporated formally into the non-random implementation. The fact that implementation was very intensive outside of the RCT district suggest that lesson design did not manifest as a type of "ownership" of the program.

Implications for PSA designs in educational evaluation. The combination of propensity scores for a treatment assigned at the school level with longitudinal developmental change in student learning spanning elementary and middle school years presented substantial methodological challenges that raise issues that should be taken up in a future study. Propensity score models in designs with clustered data tend to take one of two forms determined by the level at which the treatment is administered. One is typified by Hong and Raudenbush (2005), in which an intervention was administered to students randomly assigned within schools and the effect of the treatment on individual students was of central interest. Each school, subsequently, represented a site within the overall evaluation and the design was characterized and analyzed as a multi-site RCT. This approach has been examined extensively, most notably by Thoemmes and West (2011), who look at different forms the propensity score model might take, depending on whether assignment was independent of the clustered nature of the data. The other is typified by Stuart (2007), in which entire schools were randomized to conditions, but the effect of the treatment was assessed at the level of the school. Little guidance has been provided regarding how to conduct PSA studies when treatments have been non-randomly assigned to entire schools and in which interest in treatment impact is at the student level.

The present study added the complicating factor of observing students over elementary and middle school levels. A straightforward approach to PSA modeling in this scenario, if present-day schools followed traditional highly structured feeder patterns, would have been to estimate a model for assignment with elementary schools nested within middle schools. But owing to school choice policies in place in many districts, the relationships between middle schools and the elementary schools that feed them are not so highly

structured. A more accurate appraisal of the relationship is that they are cross-classified rather than nested. In a cross-classified model, participants are nested within two or more different types of structures, but they may be differently nested across each. Thus, in the present scenario, two students (A.1 and A.2) in elementary school A may attend separate middle schools B and C, respectively, while two students in elementary school D (D.1 and D.2) do the same. Although A.1 and A.2, and D.1 and D.2, are nested within their elementary schools, and A.1 and D.1 are nested within one middle school and A.2 and D.2 in the other, the nesting pattern breaks when spanning the elementary to middle levels.

As I explained in my discussion of the design, exploratory analyses strongly suggested that modeling middle school characteristics did not sufficiently proxy for elementary school characteristics. If the propensity score model was to account for pre-treatment experiences of study participants in equalizing across study conditions, these elementary school characteristics had to be taken into account. A proper propensity score model for this design would have incorporated the cross-classification by modeling "sequences" of elementary and middle schools as observed in the data, with these sequences cross-classified within elementary and middle schools. However, despite several failed attempts, software for estimating discrete models would not estimate a discrete model with a cross-classified design. Educational and developmental scholars that analyze the effects that instructional practices have on the long-term trajectories of student outcomes would benefit from being able to model PSA designs that require an acknowledgement of the cross-classified nature of the elementary-middle school relationship. In this study, I worked around the problem simply by modeling propensity scores at the level of student, in spite of the fact that the assignment mechanism was not at the student level.

Implications for the credibility of PSA designs. This study builds on the work of Peikes, Moreno & Orzol (2008), and Bifulco (2012) to propose a unique credibility test of the unconfoundedness assumption of the RCM in situations where PSA are applied to non-random replications of efficacious interventions. This credibility test was more rigorous than the evidence suggested by a typical balance test because it was not completely founded on the assumption that unobservable variables are conditionally randomly distributed across conditions. The credibility test was not free of assumptions and it may be limited in scope. However, provided the following conditions are met, the conclusions of the test will be definitive: *First*, the non-random replication of a treatment must take place concurrent with a continuation of the treatment to participants in the RCT; *second*, the continuation of the treatment in the RCT treatment schools must be of similar intensity such that a statistical comparison between the RCT impact and an impact calculated by a non-random method was a meaningful test of the PSA design and not a combination of the design and differences in intensity; *third*, data on all participants including those in the RCT and the replication must be available in order that such a comparison can be conducted. The assumption about the intensity of the treatment can be supported by collecting data on implementation. In this study, the data collected consisted of an interview with the Coordinator responsible for implementation in all schools during both the RCT and non-random studies, as well as limited lesson usage data.

Although the scope of such a test may be limited in that it will not apply to general PSA applications as the conditions under which the approach may be used are unique, a planful evaluator can use this as a strategic design feature to examine a localized scale-up of an efficacious treatment program.

Conclusion

Evidence from this study suggests that, consistent with the findings from the RCT, that CareerStart has potential as an evidence-based program ready to be scaled up to an effectiveness trial. However, the subject area in which CareerStart was shown to be effective was not consistent with the RCT. CareerStart has now demonstrated effects in math (in Woolley et al., in press) in an efficacy study as well as in reading (in the present study) using a credible non-randomized design. Further, this study demonstrated that although attention must still be paid to the how the components of the intervention fit to the districts in which it will be implemented, the Coordinator and principal and teacher buy-in—which I argue may be the central features of the program along with the packaged lesson—appear to be replicable. Although there remain issues related to what should be included in a scaled-up version of the program, and the feasibility of a typical district hiring a full-time Coordinator, these critical features of the design have been validated. As a consequence, continuing evaluation of CareerStart should be undertaken, with a focus on the overall effect of treatment, local effects in different contexts, and implementation fidelity. Several questions were raised by this study that could be answered with another evaluation.

This study also raised important issues about PSA models for assessing student improvement in longitudinal designs spanning elementary and middle school years, which required the use of a student-level PSA model, rather than a more-appropriate school-level model. Finally, this study proposed the statistical equivalence test used in Bifulco (2012) and Peikes, Moreno & Orzol (2008) can be applied as a design feature for evaluators looking to answer dissemination and scale-up questions in situations where an RCT treatment is continued while also disseminated non-randomly to nearby areas. This test, which requires

satisfying the assumption that the level of treatment offered to the RCT treatment participants be equivalent across the random and non-random design periods, offers clearer evidence of the credibility of the PSA than the balance test.

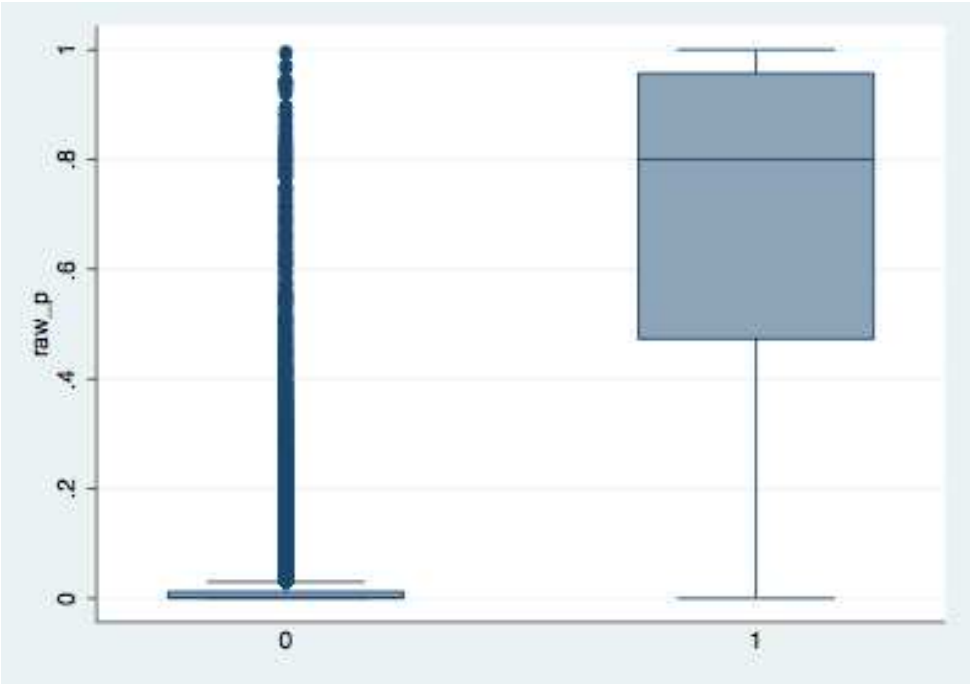

Figure 3.1. Box plot, raw propensity score.

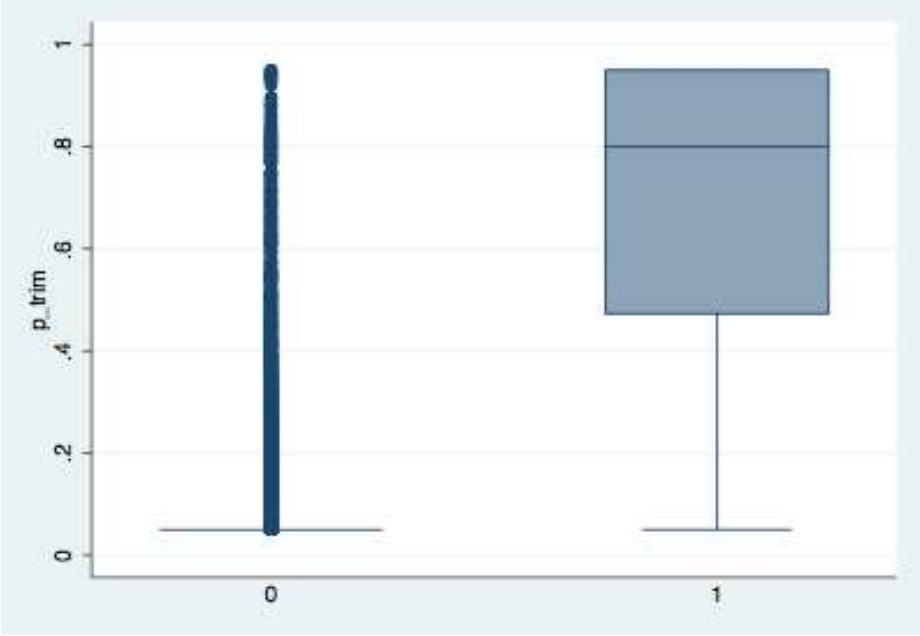

Figure 3.2. Box plot, propensity scores trimmed at .05 (comparison) and .95 (tx).

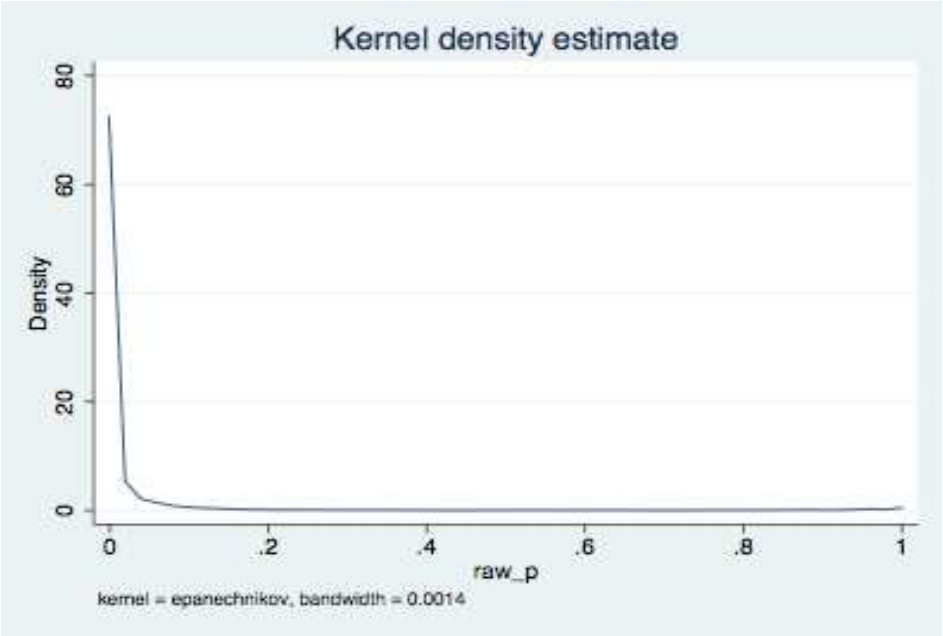

Figure 3.3. Kernel density function, raw propensities, all participants.

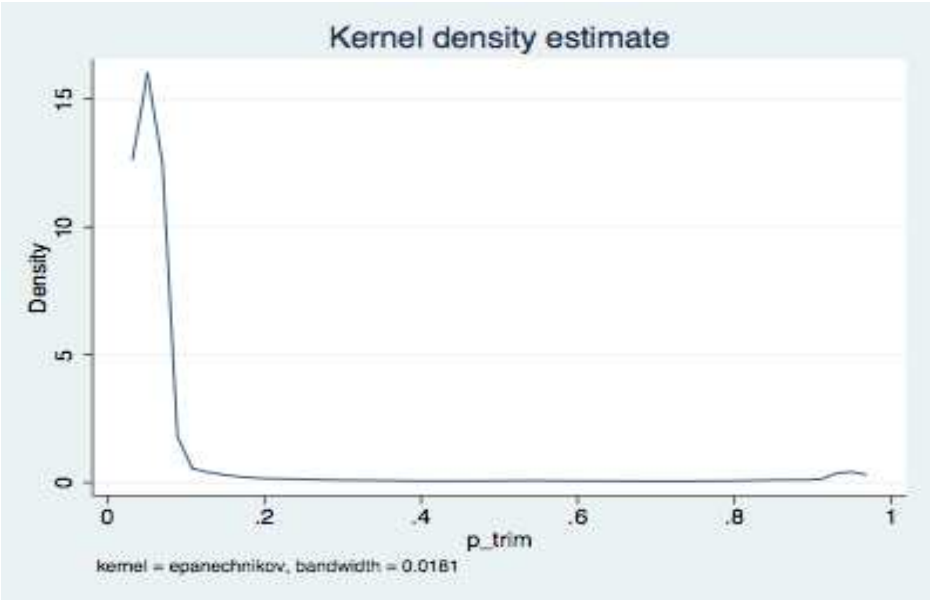

Figure 3.4. Kernel density function, trimmed propentities, all participants.

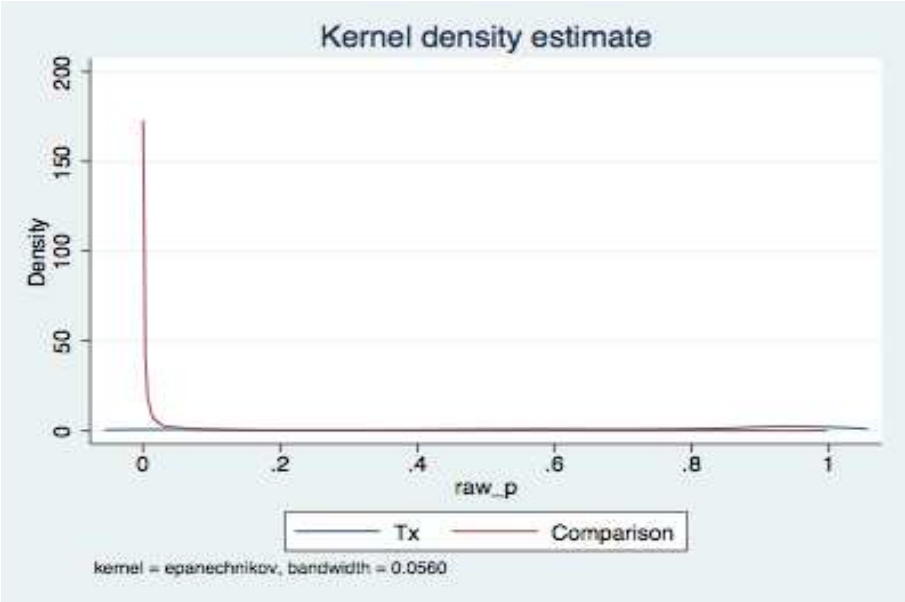

Figure 3.5. Kernel density function, raw propensities, by treatment condition.

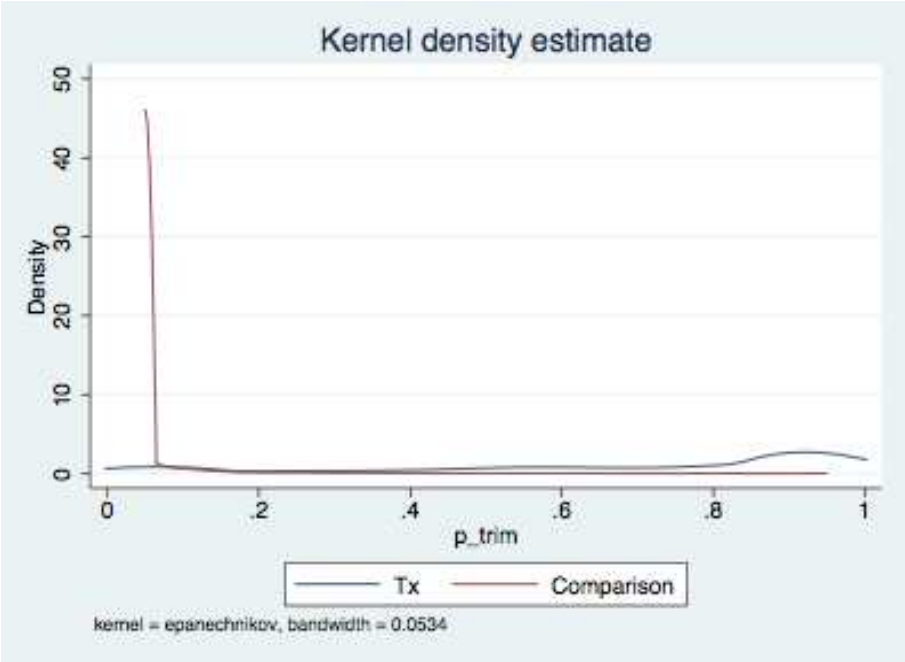

Figure 3.6. Kernel density function, trimmed propensities, by treatment condition.

Table 3.1A.

Cohort Entry and Years Observed (Aligned by Treatment Year)

	Pre-Treatment				Treatment		
Schools	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**
Group 1 (N = 13)		2005-06	2006-07	2007-08	2008-09	2009-10	2010-11
Group 2 (N = 2)		2006-07	2007-08	2008-09	2009-10	2010-11	
Group 3 (N = 3)		2007-08	2008-09	2009-10	2010-11		
Group 4 (N = 2)	2006-07	2007-08	2008-09	2009-10	2010-11		

Table 3.1B.

Cohort Entry and Grade Levels Observed (Aligned by Treatment Year)

	Pre-Treatment				Treatment		
Schools	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**
Group 1 (N = 13)		3	4	5	6	7	8
Group 2 (N = 2)		3	4	5	6	7	
Group 3 (N = 3)		3	4	5	6		
Group 4 (N = 2)	3	4	5	6	7		

Table 3.2.

Heterogeneity between Counties

	Outside RCT District		In RCT District			
N	**495243**		**11235**			
	Mean	**SD**	**Mean**	**SD**	**t value**	**p value**
Student EOG math score, 3rd grade, difference from 5th	0.01	0.62	-0.05	0.60	9.192	0.0000
Student EOG reading score, 3rd grade, difference from 5th	-0.01	0.64	0.02	0.63	-6.063	0.0000
Student EOG math score, 4th grade, difference from 5th	0.02	0.57	0.05	0.59	-6.708	0.0000
Student EOG reading score, 4th grade, difference from 5th	0.00	0.61	0.02	0.62	-3.006	0.0027
Classroom in 75th percentile academically/intellectually gifted	0.02	0.12	0.08	0.27	-24.282	0.0000
Student is Black	0.24	0.43	0.22	0.42	4.002	0.0001
No of days attended	164.30	16.27	162.39	15.92	12.381	0.0000
Classroom in 75th percentile learning disabled	0.00	0.07	0.02	0.15	-12.498	0.0000
Classroom in 75th percentile of students receiving free or reduced price l	0.07	0.26	0.17	0.38	-28.284	0.0000
Student receives free or reduced price lunch	0.37	0.48	0.36	0.48	0.585	0.5588
Student is Hispanic	0.10	0.30	0.15	0.35	-14.248	0.0000
Student has limited English proficiency	0.06	0.23	0.09	0.29	-13.164	0.0000
Student moved in current school year	0.01	0.11	0.00	0.07	11.448	0.0000
Student moved in previous school year	0.01	0.12	0.01	0.09	5.649	0.0000
Missing student school info for previous year	0.13	0.34	0.03	0.16	66.755	0.0000
Average classroom peer reading score	0.05	0.48	-0.04	0.58	16.099	0.0000
Student EOG math score, 5th grade, standardized	0.13	0.97	0.07	0.92	7.623	0.0000
Student EOG reading score, 5th grade, standardized	0.11	0.97	-0.01	0.96	12.723	0.0000
Student was limited English proficient in previous year	0.02	0.15	0.01	0.08	18.851	0.0000

Violent acts per 1,000	1.89	3.12	0.76	1.37	81.128	0.0000
Percentage Asian students	2.42	3.62	1.62	1.49	52.850	0.0000
% AYP goals met	88.18	15.09	82.79	12.36	44.775	0.0000
Percentage Black students	25.09	21.55	24.81	18.46	1.573	0.1157
School is in a mid-sized city	0.15	0.35	0.36	0.48	-46.284	0.0000
Per-pupil spending on community services	71.42	105.73	47.78	108.57	22.491	0.0000
Per pupil spending on extracurricular activity	23.05	48.14	14.90	12.42	59.161	0.0000
Per pupil spending on food services	481.48	138.50	397.15	97.33	88.435	0.0000
Pct of teachers who are fully licensed	0.97	0.03	0.98	0.02	-57.291	0.0000
Teacher is full time or receives full-time pay	0.58	0.49	0.59	0.49	-0.490	0.6245
School ABC status is high performing	0.11	0.32	0.00	0.00	251.167	0.0000
School classified as high growth	0.62	0.48	0.21	0.41	105.401	0.0000
Percentage of Hispanic students	10.80	9.86	14.97	14.14	-30.610	0.0000
School ABC status is low performing	0.26	0.44	0.30	0.46	-8.517	0.0000
School classified as no growth	0.08	0.27	0.07	0.25	4.334	0.0000
Avg composite EOG passrate (N-passed/N-taken)	0.54	0.15	0.54	0.16	2.081	0.0375
Per pupil spending on professional development	67.69	38.49	69.04	30.09	-4.592	0.0000
Avg proportion of teachers with advanced degrees	0.28	0.10	0.31	0.08	-36.890	0.0000
Per pupil spending on school leadership	497.03	153.46	452.09	145.14	31.919	0.0000
Avg proportion of National Board Certified teachers	0.11	0.08	0.08	0.05	57.933	0.0000
Per-pupil spending on student services	384.92	141.34	438.83	92.54	-59.260	0.0000
School receives Title I funds	0.66	0.47	0.26	0.44	94.249	0.0000
Total per pupil expenditures	83.90	13.19	87.57	12.59	-30.012	0.0000
Per pupil spending on transportation	282.71	118.69	296.37	16.46	-58.613	0.0000
Avg proportion of teachers that returned the following year	0.80	0.11	0.85	0.08	-54.107	0.0000
Student is American Indian	0.01	0.09	0.00	0.04	18.640	0.0000
Overage students in classroom at 75th percentile	0.00	0.05	0.01	0.11	-9.060	0.0000
Avg standardized peer EOG score	0.87	0.16	0.83	0.17	22.092	0.0000

Birthdate below standard cutoff for cohort	0.01	0.10	0.01	0.08	4.813	0.0000
Per-pupil spending on district administration	249.84	111.27	551.90	48.32	-616.404	0.0000
School is located in a large city	0.09	0.29	0.00	0.04	150.435	0.0000
Classroom standard deviation of previous year EOG math scores	0.86	0.18	0.82	0.20	22.389	0.0000
School is a magnet	0.07	0.26	0.00	0.00	194.036	0.0000
Per pupil spending on miscellaneous	0.08	1.02	0.00	0.17	37.110	0.0000
School ABC status is No Recognition	0.01	0.11	0.00	0.00	78.705	0.0000
Avg proportion of teachers that returned from previous year	0.87	0.14	0.90	0.06	-48.380	0.0000
Avg proportion of teachers that returned from previous year	0.87	0.14	0.90	0.06	0.000	0.0000
School is located in a rural area	0.52	0.50	0.19	0.39	85.410	0.0000
Avg proportion of teachers with supplemental Masters	0.20	0.09	0.21	0.08	-18.129	0.0000
Avg proportion of teachers with supplemental Masters	0.20	0.09	0.21	0.08	0.000	0.0000
Years of teacher experience	11.72	9.63	12.03	9.35	-3.295	0.0010
School has a year-round academic calendar	0.08	0.26	0.00	0.03	161.888	0.0000

Table 3.3.

Variables used in Estimation

		Imputation/Candidates for PSM	PSM	Analysis
Student Achievement or Differences in Achievement				
treatment	Non-random assignment to treatment	x	DV	x
mdiff3	Student EOG math score, 3rd grade, difference from 5th		x	
rdiff3	Student EOG reading score, 3rd grade, difference from 5th		x	
mdiff4	Student EOG math score, 4th grade, difference from 5th		x	
rdiff4	Student EOG reading score, 4th grade, difference from 5th		x	
std_ma_score_l1	4th grade standardized math achievement	x		
std_math	Student EOG math score, 5th grade, standardized	x	x	x
std_read	Student EOG reading score, 5th grade, standardized	x	x	x
Student Characteristics				
aig75	Classroom in 75th percentile academically/intellectually gifted	x	x	x
amindian		x		
asian	Student is Asian	x	x	
black	Student is Black	x	x	x
daysmem	No of days attended	x	x	x
dis75	Classroom in 75th percentile learning disabled	x	x	x
ex_aig	Student is academically/intellectually gifted	x		x
ex_dis	Student has learning disability	x		x
frl75	Classroom in 75th percentile of students receiving free or reduced price l	x	x	x
frlnch	Student receives free or reduced price lunch	x	x	x

hispanic	Student is Hispanic	x	x	x
islep	Student has limited English proficiency	x	x	x
ma_peer	Average classroom peer math score	x		x
male	Student is a male	x		x
movediy	Student moved in current school year	x	x	x
movedpy	Student moved in previous school year	x	x	x
movedpymiss	Missing student school info for previous year	x	x	
overage	Student is overage for cohort	x		
rd_peer	Average classroom peer reading score	x	x	x
waslep	Student was limited English proficient in previous year	x	x	x
oage75	Overage students in classroom at 75th percentile	x		x
peers_sd	Avg standardized peer EOG score	x		
underage	Birthdate below standard cutoff for cohort			x
Elementary School Characteristics (During 5th Grade Year for Cohort)				
actper1k	Violent acts per 1,000	x	x	
asian_mean	Percentage Asian students	x	x	
aypper	% AYP goals met	x	x	
black_mean	Percentage Black students	x	x	
city	School is in a mid-sized city	x	x	
comm_serv	Per-pupil spending on community services	x	x	
extra_cur	Per pupil spending on extracurricular activity	x	x	
food_serv	Per pupil spending on food services	x	x	
full_licen	Pct of teachers who are fully licensed	x	x	
fullpay	Teacher is full time or receives full-time pay	x	x	
high	School ABC status is high performing	x	x	
higrowth	School classified as high growth	x	x	
hispanic_mean	Percentage of Hispanic students	x	x	

low	School ABC status is low performing	x	x
nogrowth	School classified as no growth	x	x
pctfrpl	Percentage free/reduced lunch students at school	x	
perfcompeog	Avg composite EOG passrate (N-passed/N-taken)	x	x
prof_dev	Per pupil spending on professional development	x	x
sch_advdeg	Avg proportion of teachers with advanced degrees	x	x
sch_lead	Per pupil spending on school leadership	x	x
sch_nbc	Avg proportion of National Board Certified teachers	x	x
stud_serv	Per-pupil spending on student services	x	x
title1	School receives Title I funds	x	x
tot_ppx	Total per pupil expenditures	x	x
transport	Per pupil spending on transportation	x	x
willreturn	Avg proportion of teachers that returned the following year	x	x
dist_admin	Per-pupil spending on district administration	x	
largecity	School is located in a large city	x	
ma_peer_sd	Classroom standard deviation of previous year EOG math scores	x	
magnet	School is a magnet	x	
miscellan	Per pupil spending on miscellaneous	x	
nr	School ABC status is No Recognition	x	
returned	Avg proportion of teachers that returned from previous year	x	
rural	School is located in a rural area	x	
sch_dpisuppma	Avg proportion of teachers with supplemental Masters	x	
teach_exp	Years of teacher experience	x	
yrround	School has a year-round academic calendar	x	

Differences between 5th Grade and 4th Grade Elementary School Characteristics for Cohort			
d_aypper	Difference 4th-5th grades, % AYP goals met	x	x
d_full_licen	Difference, 4th-5th grade, % of teachers who are fully licensed	x	
d_hispanic_mean	Difference 4th-5th grades, % Hispanic students	x	x
d_miscellan	Difference 4th-5th grades, Per-pupil spending on miscellaneous	x	x
d_perfcompeog	Difference 4th-5th grades, school avg composite EOG passrate	x	x
d_spec_instr	Difference 4th-5th grades, per-pupil spending on special instruction	x	x
d_actper1k	Difference 4th-5th grades, violent acts per 1,000	x	
d_asian_mean	Difference 4th-5th grades, % of students who are Asian	x	
d_black_mean	Difference 4th-5th grades, % of students who are Black	x	
d_cap_outlay	Difference 4th-5th grades, per-pupil spending on capital outlays	x	
d_comm_serv	Difference 4th-5th grades, per-pupil spending on community services	x	
d_food_serv	Difference 4th-5th grades, per pupil spending on food services	x	
d_pctfrpl	Difference 4th-5th grades, % of students receiving free/reduced price lunc	x	
d_returned	Difference 4th-5th grades, proportion of teachers who returned from previo	x	
d_sch_advdeg	Difference 4th-5th grades, avg proportion of teachers with advanced degree	x	
d_sch_dpisuppma	Difference 4th-5th grades, avg proportion of teachers with supplemental Ma	x	
d_sch_lead	Difference 4th-5th grades, per pupil spending on school leadership	x	

d_sch_nbc	Difference 4th-5th grades, Avg proportion of National Board Certfified tea	x	
d_stsrate	Difference 4th-5th grades, short-term suspension rate	x	
d_stud_serv	Difference 4th-5th grades, per-pupil spending on student services	x	
d_transport	Difference 4th-5th grades, per pupil spending on transportation	x	
d_willreturn	Difference 4th-5th grades, avg proportion of teachers who return the follo	x	
Pre-Treatment Middle School Characteristics			
ms_cap_outlay	Per pupil spending on capital outlay		
ms_comm_serv	Per pupil spending on community services		
ms_extra_cur	Per pupil spending on extracurricular activities		
ms_returned	Avg proportion of teachers that returned to the school from previous year		
ms_actper1k	Violent acts per 1,000		x
ms_asian_mean	Percentage Asian students		x
ms_aypper	Percentage of AYP goals met		x
ms_exp3orlessyr	Avg. proportion of teachers with 3 or fewer years exp		x
ms_food_serv	Per pupil spending on food services		x
ms_full_licen	Pct of teachers who are fully licensed		x
ms_miscellan	Per-pupil spending on miscellaneous		
ms_perfcompeog	School avg composite EOG passrate		x
ms_prof_dev	Per pupil spending on professional development		x
ms_sch_advdeg	Avg proportion of teachers with advanced degrees		x
ms_sch_dpisuppma	Avg proportion of teachers with supplemental Masters degrees		x
ms_sch_lead	Per pupil spending on school leadership		x

ms_sch_nbc	Avg proportion of National Board Certified teachers		x	
ms_stsrate	Short-term suspension rate		x	
ms_willreturn	Avg proportion of teachers that return to the school in following year		x	
Post-Treatment Middle School Characteristics				
a_pctfrpl	Percent free/reduced price lunch	x		x
a_title1	School has Title I status	x		x
a_actper1k	Violent acts per 1,000	x		x
a_amindian_mean	Percentage American Indian students	x		
a_asian_mean	Percentage Asian students	x		
a_aypper	% AYP goals met	x		
a_black_mean	Percentage Black students	x		x
a_cap_outlay	Per-pupil spending on capital outlays	x		
a_comm_serv	Per-pupil spending on community services	x		
a_exp3orlessyr	Avg. proportion of teachers with 3 or fewer years exp	x		x
a_extra_cur	Per-pupil spending on extracurricular activities	x		
a_food_serv	Per-pupil spending on food services	x		
a_full_licen	Pct of teachers who are fully licensed	x		x
a_hispanic_mean	Percentage Hispanic students	x		x
a_instr_sup	Per pupil spending on instructional support	x		
a_miscellan	Per-pupil spending on miscellaneous	x		
a_n_years_open	Number of years the school was open between 2005-2011	x		
a_prof_dev	Per-pupil spending on professional development	x		
a_propteachers	Teacher-student ratio			
a_reg_instr	Per pupil-spending on regular instruction	x		x
a_returned	Avg proportion of teachers that returned to the school from previous year	x		x
a_sch_advdeg	Avg proportion of teachers with advanced degrees	x		x

a_sch_dpisuppma	Avg proportion of teachers with supplemental Masters degrees	x	x
a_sch_lead	Per-pupil spending on school leadership	x	
a_sch_maint	Per-pupil spending on school maintenance	x	
a_sch_nbc	Avg proportion of National Board Certified teachers	x	x
a_spec_instr	Per pupil-spending on special instruction	x	
a_stsrate	Short-term suspension rate	x	x
a_stud_serv	Per pupil-spending on student services	x	
a_tot_distcap	Total district spending on capital outlay	x	
a_willreturn	Avg proportion of teachers that return to the school the following year	x	
a_tcount	# of clssrm tchrs according to 2010 pay data *see codebook for other yrs	x	
a_adm	Average daily membership		x
a_schoolppe	Per-pupil spending		x
magnet	School is a magnet school		x
low	School ABC status is low performing		x
high	School ABC status is high performing		x
nogrowth	School classified as no growth		x
higrowth	School classified as high growth		x
magnet	School is a magnet		x
sch_dpisuppma	Avg proportion of teachers with supplemental Masters		x

Table 3.4.

Univariate Statistics

	Mean	**SD**	**Min**	**Median**	**Max**	**N(Miss)**
Student EOG math score, 3rd grade, difference from 5th	0.00	0.63	-4.23	-0.01	3.91	19185
Student EOG reading score, 3rd grade, difference from 5th	-0.02	0.65	-4.43	-0.02	3.95	19528
Student EOG math score, 4th grade, difference from 5th	0.01	0.58	-4.01	0.01	3.48	10719
Student EOG reading score, 4th grade, difference from 5th	0.00	0.62	-3.85	0.00	4.35	11092
Classroom in 75th percentile academically/intellectually gifted	0.01	0.12	0.00	0.00	1.00	7730
Student is Asian	0.03	0.16	0.00	0.00	1.00	0
Student is Black	0.27	0.45	0.00	0.00	1.00	0
No of days attended	162.23	21.30	0.00	166.00	186.00	8092
Classroom in 75th percentile learning disabled	0.01	0.08	0.00	0.00	1.00	7730
Classroom in 75th percentile of students receiving free or reduced price l	0.08	0.28	0.00	0.00	1.00	7730
Student receives free or reduced price lunch	0.41	0.49	0.00	0.00	1.00	73
Student is Hispanic	0.10	0.31	0.00	0.00	1.00	0
Student has limited English proficiency	0.06	0.24	0.00	0.00	1.00	8
Student moved in current school year	0.04	0.20	0.00	0.00	1.00	0
Student moved in previous school year	0.05	0.22	0.00	0.00	1.00	0
Missing student school info for previous year	0.14	0.35	0.00	0.00	1.00	7730
Average classroom peer reading score	0.03	0.49	-3.17	0.03	1.76	12956
Student EOG math score, 5th grade, standardized	0.06	0.98	-2.97	0.06	2.57	8129
Student EOG reading score, 5th grade, standardized	0.04	0.99	-2.97	0.11	2.81	8459
Student was limited English proficient in previous year	0.02	0.14	0.00	0.00	1.00	8
Violent acts per 1,000	1.97	3.38	0.00	0.00	227.27	7730
Percentage Asian students	2.47	3.66	0.00	1.27	47.97	7730

% AYP goals met	86.93	16.58	0.00	92.30	100.00	7739
Percentage Black students	26.59	22.32	0.00	20.68	99.19	7730
School is in a mid-sized city	0.16	0.37	0.00	0.00	1.00	7891
Per-pupil spending on community services	72.47	106.48	-12.96	14.98	984.12	7730
Per pupil spending on extracurricular activity	23.04	46.73	-4.23	8.27	586.22	7730
Per pupil spending on food services	479.94	140.88	14.98	476.00	3570.73	7730
Pct of teachers who are fully licensed	0.97	0.03	0.71	0.98	1.00	7730
Teacher is full time or receives full-time pay	0.54	0.50	0.00	1.00	1.00	13008
School ABC status is high performing	0.14	0.35	0.00	0.00	1.00	7764
School classified as high growth	0.68	0.47	0.00	1.00	1.00	7751
Percentage of Hispanic students	11.16	10.16	0.00	7.56	72.50	7730
School ABC status is low performing	0.31	0.46	0.00	0.00	1.00	7764
School classified as no growth	0.06	0.25	0.00	0.00	1.00	7751
Avg composite EOG passrate (N-passed/N-taken)	0.53	0.16	0.00	0.54	1.00	7730
Per pupil spending on professional development	68.27	39.70	11.42	59.14	1319.40	7730
Avg proportion of teachers with advanced degrees	0.28	0.10	0.00	0.27	0.65	7730
Per pupil spending on school leadership	500.15	160.32	8.31	465.33	5936.89	7730
Avg proportion of National Board Certified teachers	0.10	0.08	0.00	0.09	0.50	7730
Per-pupil spending on student services	386.02	144.81	38.12	361.55	4217.15	7730
School receives Title I funds	0.65	0.48	0.00	1.00	1.00	7730
Total per pupil expenditures	84.34	13.95	59.48	81.63	622.62	7730
Per pupil spending on transportation	285.49	117.96	17.20	257.53	2334.39	7730
Avg proportion of teachers that returned the following year	0.80	0.11	0.00	0.82	1.00	8452
Difference 4th-5th grades, % AYP goals met	-4.70	14.69	-99.06	0.00	50.00	9137
Difference 4th-5th grades, % Hispanic students	0.61	2.26	-10.48	0.49	38.40	9130
Difference 4th-5th grades, Per-pupil spending on miscellaneous	-59.07	36.03	-635.71	-48.03	1.50	9130
Difference 4th-5th grades, school avg composite EOG passrate	-0.13	0.07	-0.47	-0.13	0.26	9130

Difference 4th-5th grades, per-pupil spending on special instruction	111.28	262.18	-8019.55	103.10	8265.43	9130
Violent acts per 1,000	10.20	7.82	0.00	8.86	47.62	0
Percentage Asian students	2.26	2.83	0.00	1.29	18.70	0
Percentage of AYP goals met	86.50	10.92	52.00	88.20	100.00	0
Avg. proportion of teachers with 3 or fewer years exp	0.24	0.10	0.00	0.23	0.55	0
Per pupil spending on food services	427.58	133.41	-0.21	427.04	1131.40	0
Pct of teachers who are fully licensed	0.91	0.08	0.53	0.93	1.00	0
Per-pupil spending on miscellaneous	58.40	35.70	14.75	46.62	237.43	0
School avg composite EOG passrate	0.64	0.14	0.18	0.64	1.00	0
Per pupil spending on professional development	67.91	32.12	20.85	57.71	257.64	0
Avg proportion of teachers with advanced degrees	0.27	0.09	0.04	0.27	0.52	0
Avg proportion of teachers with supplemental Masters degrees	0.19	0.07	0.02	0.19	0.42	0
Per pupil spending on school leadership	524.25	146.63	239.69	496.71	1523.45	0
Avg proportion of National Board Certified teachers	0.08	0.06	0.00	0.06	0.35	0
Short-term suspension rate	33.20	23.96	0.62	27.66	199.66	0
Avg proportion of teachers that return to the school in following year	0.75	0.10	0.43	0.76	1.00	0
Student is American Indian	0.01	0.09	0.00	0.00	1.00	0
Overage students in classroom at 75th percentile	0.00	0.06	0.00	0.00	1.00	7730
Avg standardized peer EOG score	0.87	0.16	0.08	0.88	2.53	7895
Birthdate below standard cutoff for cohort	0.01	0.11	0.00	0.00	1.00	8
Per-pupil spending on district administration	258.51	119.30	90.62	223.48	1194.23	7730
School is located in a large city	0.10	0.30	0.00	0.00	1.00	7891
Classroom standard deviation of previous year EOG math scores	0.86	0.18	0.00	0.87	2.84	7898
School is a magnet	0.06	0.24	0.00	0.00	1.00	7858
Per pupil spending on miscellaneous	0.07	0.95	-5.08	0.00	11.38	7730

School ABC status is No Recognition	0.02	0.13	0.00	0.00	1.00	7764
Avg proportion of teachers that returned from previous year	0.87	0.14	0.00	0.90	1.00	7732
School is located in a rural area	0.50	0.50	0.00	0.00	1.00	7891
Avg proportion of teachers with supplemental Masters	0.20	0.09	0.00	0.19	0.59	7730
Years of teacher experience	11.62	9.62	0.00	9.00	49.00	14931
School has a year-round academic calendar	0.08	0.26	0.00	0.00	1.00	7730
Difference 4th-5th grades, violent acts per 1,000	0.32	3.48	-58.14	0.00	69.44	9130
Difference 4th-5th grades, % of students who are Asian	0.05	0.92	-6.08	0.00	5.80	9130
Difference 4th-5th grades, % of students who are Black	-0.10	3.75	-52.24	-0.07	26.12	9130
Difference 4th-5th grades, per-pupil spending on capital outlays	-120.93	2190.73	-25754.72	0.00	18694.10	9130
Difference 4th-5th grades, per-pupil spending on community services	23.72	91.24	-479.37	0.00	869.76	9130
Difference 4th-5th grades, per pupil spending on food services	29.73	85.16	-809.75	38.43	1510.15	9130
Difference 4th-5th grades, % of students receiving free/reduced price lunc	0.13	7.58	-99.00	0.40	61.30	9130
Difference 4th-5th grades, proportion of teachers who returned from previo	0.00	0.14	-0.59	-0.01	0.95	9130
Difference 4th-5th grades, avg proportion of teachers with advanced degree	0.01	0.05	-0.34	0.01	0.30	9130
Difference 4th-5th grades, avg proportion of teachers with supplemental Ma	0.00	0.05	-0.34	0.00	0.30	9130
Difference 4th-5th grades, per pupil spending on school leadership	27.11	88.33	-2238.75	23.26	1140.24	9130
Difference 4th-5th grades, Avg proportion of National Board Certfified tea	0.01	0.04	-0.22	0.00	0.22	9130
Difference 4th-5th grades, short-term suspension rate	0.13	6.05	-168.46	0.00	248.38	9130
Difference 4th-5th grades, per-pupil spending on student services	-45.68	121.16	-8031.95	-33.97	860.39	9130
Difference 4th-5th grades, per pupil spending on	30.22	37.00	-1653.42	26.03	446.73	9130

transportation						
Difference 4th-5th grades, avg proportion of teachers who return the follo	0.02	0.13	-0.86	0.02	0.75	10420
Per pupil spending on capital outlay	344.39	1614.61	-12.51	26.87	19989.74	0
Per pupil spending on community services	6.64	17.41	-0.03	0.00	130.77	0
Per pupil spending on extracurricular activities	33.54	36.31	0.00	29.87	283.10	0
Avg proportion of teachers that returned to the school from previous year	0.87	0.14	0.00	0.90	1.00	0

Table 3.5.

Model Variations

		Robustness Variations				
		Not Imputed	**Group 4**	**No Group 3**	**No Group 3 or 2**	**No Movers**
Label	**-**	**R1**	**R2**	**R3**	**R4**	**R5**
Group 4[1]	No	No	*Yes*	No	No	No
Group 3[2]	Yes	Yes	Yes	*No*	*No*	Yes
Group 2[3]	Yes	Yes	Yes	Yes	*No*	Yes
Unconditional Q1-Q2	Yes	Yes	*No*	*No*	*No*	Yes
Condition Q1-Q2	Yes	Yes	Yes	Yes	Yes	Yes
Unconditional Q3D[4]	Yes	Yes	*No*	*No*	*No*	Yes
Unconditional Q3S[4]	Yes	Yes	*No*	*No*	*No*	Yes
Conditional Q3D[4]	Yes	Yes	Yes	Yes	Yes	Yes
Conditional Q3S[4]	Yes	Yes	Yes	Yes	Yes	Yes
Imputation	No	*Yes*	No	No	No	No
Interactions in PSM	No	No	No	No	No	No
Students moving during treatment	Yes	Yes	No	No	No	*No*

1. Group 4 started in 7th grade and have only one year of MS tx and data by 2010-11.
2. 2. Group 3 only had one year of MS tx and data by 2010-11.
3. Group 2 had two years of MS tx and data by 2010-11.
4. Q3D is the interaction model that tests the difference between RCT-control and nonRCT.

Q3S is the status model that compares RCT-control and non-RCT separately with the comparisons.

Table 3.6.

Correlations between Outcome Proxies and Candidates for PSM Covariates

	Pearson Correlation	
	5th Grade Math	**5th Grade Reading**
Student EOG math score, 3rd grade, difference from 5th	0.341	0.088
Student EOG reading score, 3rd grade, difference from 5th	0.149	0.370
Student EOG math score, 4th grade, difference from 5th	0.280	0.030
Student EOG reading score, 4th grade, difference from 5th	0.090	0.328
Classroom in 75th percentile academically/intellectually gifted	0.151	0.150
Student is Asian	0.093	0.048
Student is Black	-0.319	-0.288
No of days attended	0.076	0.069
Classroom in 75th percentile learning disabled	-0.055	-0.064
Classroom in 75th percentile of students receiving free or reduced price lunch	-0.180	-0.186
Student receives free or reduced price lunch	-0.354	-0.376
Student is Hispanic	-0.087	-0.143
Student has limited English proficiency	-0.151	-0.214
Student moved in current school year	-0.077	-0.077
Student moved in previous school year	-0.086	-0.085
Missing student school info for previous year	-0.029	-0.025
Average classroom peer reading score	0.382	0.393
Student EOG math score, 5th grade, standardized	1.000	0.729
Student EOG reading score, 5th grade, standardized	0.729	1.000
Student was limited English proficient in previous year	0.045	0.020
Violent acts per 1,000	-0.084	-0.070
Percentage Asian students	0.118	0.114
% AYP goals met	0.188	0.177
Percentage Black students	-0.216	-0.209
School is in a mid-sized city	-0.058	-0.052
Per-pupil spending on community services	0.067	0.074
Per pupil spending on extracurricular activity	0.035	0.022
Per pupil spending on food services	-0.200	-0.201
Pct of teachers who are fully licensed	0.106	0.105
Teacher is full time or receives full-time pay	0.363	0.379
School ABC status is high performing	0.227	0.224
School classified as high growth	0.158	0.101
Percentage of Hispanic students	-0.104	-0.132

School ABC status is low performing	-0.276	-0.264
School classified as no growth	-0.135	-0.088
Avg composite EOG passrate (N-passed/N-taken)	0.345	0.347
Per pupil spending on professional development	-0.167	-0.163
Avg proportion of teachers with advanced degrees	0.067	0.071
Per pupil spending on school leadership	-0.059	-0.057
Avg proportion of National Board Certified teachers	0.123	0.123
Per-pupil spending on student services	-0.072	-0.069
School receives Title I funds	-0.202	-0.201
Total per pupil expenditures	-0.122	-0.131
Per pupil spending on transportation	-0.026	-0.015
Avg proportion of teachers that returned the following year	0.077	0.078
Difference 4th-5th grades, % AYP goals met	0.048	0.046
Difference 4th-5th grades, % Hispanic students	-0.053	-0.053
Difference 4th-5th grades, Per-pupil spending on miscellaneous	0.050	0.061
Difference 4th-5th grades, school avg composite EOG passrate	0.080	0.117
Difference 4th-5th grades, per-pupil spending on special instruction	0.074	0.056
Violent acts per 1,000	-0.085	-0.090
Percentage Asian students	0.114	0.106
Percentage of AYP goals met	0.193	0.191
Avg. proportion of teachers with 3 or fewer years exp	-0.086	-0.089
Per pupil spending on food services	-0.139	-0.151
Percentage of teachers who are fully licensed	0.161	0.153
Per-pupil spending on miscellaneous	-0.044	-0.059
School avg composite EOG passrate	0.296	0.290
Per pupil spending on professional development	-0.142	-0.150
Avg proportion of teachers with advanced degrees	0.107	0.109
Avg proportion of teachers with supplemental Masters degrees	0.050	0.054
Per pupil spending on school leadership	-0.045	-0.045
Avg proportion of National Board Certified teachers	0.147	0.146
Short-term suspension rate	-0.194	-0.188
Avg proportion of teachers that return to the school in following year	0.095	0.098
Student is American Indian	-0.024	-0.027
Overage students in classroom at 75th percentile	-0.053	-0.064
Avg standardized peer EOG score	-0.044	-0.049
Birthdate below standard cutoff for cohort	0.038	0.034
Per-pupil spending on district administration	-0.014	-0.024
School is located in a large city	0.016	0.013
Classroom standard deviation of previous year EOG math scores	0.005	0.006

School is a magnet	0.014	0.023
Per pupil spending on miscellaneous	-0.011	-0.006
School ABC status is No Recognition	-0.019	0.006
Avg proportion of teachers that returned from previous year	0.086	0.095
School is located in a rural area	-0.025	-0.018
Avg proportion of teachers with supplemental Masters	0.023	0.027
Years of teacher experience	0.032	0.039
School has a year-round academic calendar	0.054	0.051
Difference 4th-5th grades, violent acts per 1,000	-0.028	-0.015
Difference 4th-5th grades, % of students who are Asian	0.035	0.036
Difference 4th-5th grades, % of students who are Black	0.017	0.027
Difference 4th-5th grades, per-pupil spending on capital outlays	-0.022	-0.009
Difference 4th-5th grades, per-pupil spending on community services	0.043	0.051
Difference 4th-5th grades, per pupil spending on food services	-0.032	-0.022
Difference 4th-5th grades, % of students receiving free/reduced price lunches	-0.041	-0.047
Difference 4th-5th grades, proportion of teachers who returned from previous year	0.025	0.032
Difference 4th-5th grades, avg proportion of teachers with advanced degrees	-0.011	-0.013
Difference 4th-5th grades, avg proportion of teachers with supplemental Masters degrees	-0.013	-0.015
Difference 4th-5th grades, per pupil spending on school leadership	-0.042	-0.042
Difference 4th-5th grades, Avg proportion of National Board Certfified teachers	0.013	0.023
Difference 4th-5th grades, short-term suspension rate	-0.022	-0.012
Difference 4th-5th grades, per-pupil spending on student services	-0.026	-0.031
Difference 4th-5th grades, per pupil spending on transportation	-0.015	-0.012
Difference 4th-5th grades, avg proportion of teachers who return the following year	-0.025	-0.027
Per pupil spending on capital outlay	0.019	0.018
Per pupil spending on community services	0.032	0.031
Per pupil spending on extracurricular activities	-0.013	-0.001
Avg proportion of teachers that returned to the school from previous year	-0.020	-0.022

Table 3.7.

Propensity Score Model—Imputed Data (M = 5)

	Est.	SE	t value	p value	
Intercept	-14.829	1.282	-11.566	0.000	***
Student EOG math score, 3rd grade, difference from 5th	-0.156	0.051	-3.050	0.002	**
Student EOG reading score, 3rd grade, difference from 5th	0.185	0.047	3.897	0.000	***
Student EOG math score, 4th grade, difference from 5th	0.206	0.054	3.828	0.000	***
Student EOG reading score, 4th grade, difference from 5th	-0.032	0.049	-0.648	0.517	
Classroom in 75th percentile academically/intellectually gifted	2.660	0.149	17.831	0.000	***
Student is Asian	-0.388	0.209	-1.855	0.064	
Student is Black	-0.142	0.076	-1.874	0.061	
No of days attended	-0.001	0.002	-0.695	0.487	
Classroom in 75th percentile learning disabled	0.696	0.289	2.413	0.016	*
Classroom in 75th percentile of students receiving free or reduced price l	0.481	0.091	5.302	0.000	***
Student receives free or reduced price lunch	-0.057	0.080	-0.715	0.475	
Student is Hispanic	0.007	0.118	0.061	0.951	
Student has limited English proficiency	-0.049	0.139	-0.357	0.721	
Student moved in current school year	-0.237	0.188	-1.257	0.209	
Student moved in previous school year	-0.506	0.133	-3.806	0.000	***
Missing student school info for previous year	-3.334	0.208	-16.026	0.000	***
Average classroom peer reading score	-0.623	0.077	-8.111	0.000	***
Student EOG math score, 5th grade, standardized	0.086	0.046	1.842	0.065	
Student EOG reading score, 5th grade, standardized	-0.091	0.047	-1.945	0.052	
Student was limited English proficient in previous year	-0.303	0.201	-1.511	0.131	
Violent acts per 1,000	-0.311	0.017	-17.748	0.000	***
Percentage Asian students	-0.111	0.016	-7.008	0.000	***
% AYP goals met	-0.080	0.003	-25.496	0.000	***

Percentage Black students	-0.037	0.002	-15.575	0.000	***
School is in a mid-sized city	2.121	0.073	29.064	0.000	***
Per-pupil spending on community services	-0.003	0.000	-12.704	0.000	***
Per pupil spending on extracurricular activity	-0.016	0.001	-15.460	0.000	***
Per pupil spending on food services	-0.005	0.000	-13.326	0.000	***
Pct of teachers who are fully licensed	13.558	1.028	13.191	0.000	***
Teacher is full time or receives full-time pay	-0.069	0.079	-0.872	0.383	
School ABC status is high performing	0.829	0.105	7.883	0.000	***
School classified as high growth	0.195	0.066	2.941	0.003	**
Percentage of Hispanic students	0.038	0.003	11.596	0.000	***
School ABC status is low performing	-0.556	0.117	-4.738	0.000	***
School classified as no growth	-0.350	0.103	-3.404	0.001	***
Avg composite EOG passrate (N-passed/N-taken)	-0.219	0.625	-0.350	0.726	
Per pupil spending on professional development	0.002	0.001	2.645	0.008	**
Avg proportion of teachers with advanced degrees	3.990	0.316	12.617	0.000	***
Per pupil spending on school leadership	-0.006	0.000	-21.116	0.000	***
Avg proportion of National Board Certified teachers	-8.249	0.424	-19.467	0.000	***
Per-pupil spending on student services	-0.003	0.000	-11.218	0.000	***
School receives Title I funds	-3.084	0.093	-33.110	0.000	***
Total per pupil expenditures	0.160	0.005	34.437	0.000	***
Per pupil spending on transportation	-0.003	0.000	-7.960	0.000	***
Avg proportion of teachers that returned the following year	1.277	0.292	4.372	0.000	***
Difference 4th-5th grades, % AYP goals met	0.002	0.003	0.595	0.552	
Difference 4th-5th grades, % Hispanic students	0.050	0.012	4.025	0.000	***
Difference 4th-5th grades, Per-pupil spending on miscellaneous	0.026	0.001	17.795	0.000	***
Difference 4th-5th grades, school avg composite EOG passrate	0.132	0.610	0.217	0.828	
Difference 4th-5th grades, per-pupil spending on special instruction	-0.002	0.000	-14.345	0.000	***

Violent acts per 1,000	-0.033	0.004	-8.723	0.000	***
Percentage Asian students	-0.066	0.020	-3.327	0.001	***
Percentage of AYP goals met	-0.059	0.004	-14.516	0.000	***
Avg. proportion of teachers with 3 or fewer years exp	-6.212	0.423	-14.670	0.000	***
Per pupil spending on food services	-0.003	0.000	-12.200	0.000	***
Pct of teachers who are fully licensed	8.762	0.648	13.518	0.000	***
Per-pupil spending on miscellaneous	0.015	0.001	12.428	0.000	***
School avg composite EOG passrate	-0.917	0.505	-1.817	0.069	
Per pupil spending on professional development	0.022	0.001	24.018	0.000	***
Avg proportion of teachers with advanced degrees	-8.459	0.714	-11.839	0.000	***
Avg proportion of teachers with supplemental Masters degrees	16.512	0.895	18.445	0.000	***
Per pupil spending on school leadership	0.001	0.000	3.415	0.001	***
Avg proportion of National Board Certified teachers	-20.525	0.775	-26.489	0.000	***
Short-term suspension rate	-0.021	0.002	-10.924	0.000	***
Avg proportion of teachers that return to the school in following year	0.720	0.383	1.880	0.060	

Table 3.8.

Balance before PSA Model (Variables in PSA Model)

	Inverse Probability Weighted Means			
	Comparison	**Treatment**	**Difference (d)**	**d > |.25|**
N per group	81217	6184		
Student EOG math score, 3rd grade, difference from 5th	0.00	-0.01	-0.01	
Student EOG reading score, 3rd grade, difference from 5th	-0.03	0.03	0.07	
Student EOG math score, 4th grade, difference from 5th	0.01	0.02	0.01	
Student EOG reading score, 4th grade, difference from 5th	0.00	0.02	0.03	
Classroom in 75th percentile academically/intellectually gifted	0.01	0.07	0.23	
Student is Asian	0.03	0.01	-0.06	
Student is Black	0.28	0.23	-0.09	
No of days attended	161.78	160.60	-0.04	
Classroom in 75th percentile learning disabled	0.01	0.02	0.11	
Classroom in 75th percentile of students receiving free or reduced price l	0.08	0.19	0.23	
Student receives free or reduced price lunch	0.40	0.44	0.05	
Student is Hispanic	0.10	0.16	0.12	
Student has limited English proficiency	0.06	0.09	0.09	
Student moved in current school year	0.04	0.04	-0.02	
Student moved in previous school year	0.05	0.04	-0.03	
Missing student school info for previous year	0.15	0.04	-0.26	*
Average classroom peer reading score	0.02	-0.03	-0.07	
Student EOG math score, 5th grade, standardized	0.03	0.03	0.00	
Student EOG reading score, 5th grade, standardized	0.01	-0.02	-0.02	

Student was limited English proficient in previous year	0.02	0.02	-0.01	
Violent acts per 1,000	2.06	1.11	-0.24	
Percentage Asian students	2.54	1.68	-0.20	
% AYP goals met	88.50	64.89	-0.63	*
Percentage Black students	27.29	22.42	-0.16	
School is in a mid-sized city	0.15	0.33	0.30	*
Per-pupil spending on community services	72.94	69.56	-0.02	
Per pupil spending on extracurricular activity	24.02	11.64	-0.24	
Per pupil spending on food services	480.74	469.37	-0.06	
Pct of teachers who are fully licensed	0.97	0.98	0.28	*
Teacher is full time or receives full-time pay	0.54	0.50	-0.05	
School ABC status is high performing	0.13	0.20	0.13	
School classified as high growth	0.68	0.66	-0.02	
Percentage of Hispanic students	10.80	16.41	0.31	*
School ABC status is low performing	0.31	0.31	0.00	
School classified as no growth	0.06	0.07	0.02	
Avg composite EOG passrate (N-passed/N-taken)	0.53	0.55	0.08	
Per pupil spending on professional development	68.30	71.57	0.05	
Avg proportion of teachers with advanced degrees	0.28	0.32	0.30	*
Per pupil spending on school leadership	499.40	514.13	0.06	
Avg proportion of National Board Certified teachers	0.10	0.10	-0.07	
Per-pupil spending on student services	381.53	441.41	0.31	*
School receives Title I funds	0.66	0.50	-0.23	
Total per pupil expenditures	83.89	91.10	0.37	*
Per pupil spending on transportation	284.28	306.95	0.14	
Avg proportion of teachers that returned the following year	0.80	0.84	0.26	*
Difference 4th-5th grades, % AYP goals met	-3.90	-16.05	-0.37	*

Difference 4th-5th grades, % Hispanic students	0.55	1.17	0.19	
Difference 4th-5th grades, Per-pupil spending on miscellaneous	-60.56	-37.35	0.48	*
Difference 4th-5th grades, school avg composite EOG passrate	-0.14	-0.09	0.35	*
Difference 4th-5th grades, per-pupil spending on special instruction	117.17	115.71	0.00	
Violent acts per 1,000	10.25	9.50	-0.07	
Percentage Asian students	2.32	1.47	-0.27	*
Percentage of AYP goals met	86.84	82.00	-0.29	*
Avg. proportion of teachers with 3 or fewer years exp	0.24	0.19	-0.43	*
Per pupil spending on food services	427.98	422.26	-0.03	
Pct of teachers who are fully licensed	0.91	0.94	0.31	*
Per-pupil spending on miscellaneous	58.13	61.94	0.07	
School avg composite EOG passrate	0.64	0.61	-0.13	
Per pupil spending on professional development	67.49	73.42	0.13	
Avg proportion of teachers with advanced degrees	0.27	0.30	0.32	*
Avg proportion of teachers with supplemental Masters degrees	0.19	0.22	0.30	*
Per pupil spending on school leadership	520.98	567.17	0.21	
Avg proportion of National Board Certified teachers	0.08	0.07	-0.14	
Short-term suspension rate	33.47	29.59	-0.13	
Avg proportion of teachers that return to the school in following year	0.75	0.77	0.20	

Table 3.9.

Balance before PSA Model (Variables Not in PSA Model—Supplement)

	Inverse Probability Weighted Means			
	Comparison	**Treatment**	**Difference (d)**	**d > |.25|**
N per group	81217	6184		
Student is American Indian	0.01	0.00	-0.06	
Overage students in classroom at 75th percentile	0.00	0.01	0.08	
Avg standardized peer EOG score	0.87	0.85	-0.10	
Birthdate below standard cutoff for cohort	0.01	0.01	-0.03	
Per-pupil spending on district administration	240.76	492.36	1.63	*
School is located in a large city	0.11	0.00	-0.34	*
Classroom standard deviation of previous year EOG math scores	0.87	0.83	-0.13	
School is a magnet	0.06	0.02	-0.15	
Per pupil spending on miscellaneous	0.08	-0.02	-0.09	
School ABC status is No Recognition	0.02	0.03	0.05	
Avg proportion of teachers that returned from previous year	0.86	0.91	0.26	*
School is located in a rural area	0.50	0.38	-0.17	
Avg proportion of teachers with supplemental Masters	0.20	0.24	0.34	*
Years of teacher experience	11.42	13.21	0.13	
School has a year-round academic calendar	0.08	0.01	-0.22	
Difference 4th-5th grades, violent acts per 1,000	0.35	-0.10	-0.11	
Difference 4th-5th grades, % of students who are Asian	0.07	-0.06	-0.10	
Difference 4th-5th grades, % of students who are Black	-0.12	-0.59	-0.10	
Difference 4th-5th grades, per-pupil spending on capital outlays	-185.61	-160.38	0.01	
Difference 4th-5th grades, per-pupil spending on community services	23.35	27.90	0.03	
Difference 4th-5th grades, per	26.60	41.73	0.15	

pupil spending on food services				
Difference 4th-5th grades, % of students receiving free/reduced price lunc	0.19	-1.13	-0.08	
Difference 4th-5th grades, proportion of teachers who returned from previo	-0.01	0.00	0.05	
Difference 4th-5th grades, avg proportion of teachers with advanced degree	0.01	0.02	0.14	
Difference 4th-5th grades, avg proportion of teachers with supplemental Ma	0.00	0.01	0.17	
Difference 4th-5th grades, per pupil spending on school leadership	24.64	30.87	0.05	
Difference 4th-5th grades, Avg proportion of National Board Certfified tea	0.01	0.00	-0.19	
Difference 4th-5th grades, short-term suspension rate	0.15	0.22	0.01	
Difference 4th-5th grades, per-pupil spending on student services	-47.92	-8.99	0.25	
Difference 4th-5th grades, per pupil spending on transportation	30.91	15.51	-0.34	*
Difference 4th-5th grades, avg proportion of teachers who return the follo	0.03	0.02	-0.07	
Per pupil spending on capital outlay	366.99	47.56	-0.19	
Per pupil spending on community services	6.60	7.17	0.02	
Per pupil spending on extracurricular activities	31.06	66.05	0.58	*
Avg proportion of teachers that returned to the school from previous year	0.87	0.85	-0.11	

Table 3.10.

Balance after PSA Model (Variables in PSA Model)

	Inverse Probability Weighted Means			
	Comparison	**Treatment**	**Difference (d)**	**d > |.25|**
N per group	81217	6184		
Student EOG math score, 3rd grade, difference from 5th	0.00	-0.02	-0.01	
Student EOG reading score, 3rd grade, difference from 5th	-0.03	0.01	0.03	
Student EOG math score, 4th grade, difference from 5th	0.01	0.02	0.01	
Student EOG reading score, 4th grade, difference from 5th	-0.01	-0.01	0.00	
Classroom in 75th percentile academically/intellectually gifted	0.01	0.03	0.07	
Student is Asian	0.02	0.01	-0.06	
Student is Black	0.27	0.22	-0.07	
No of days attended	163.99	164.43	0.01	
Classroom in 75th percentile learning disabled	0.00	0.01	0.02	
Classroom in 75th percentile of students receiving free or reduced price l	0.08	0.13	0.07	
Student receives free or reduced price lunch	0.40	0.43	0.04	
Student is Hispanic	0.09	0.11	0.03	
Student has limited English proficiency	0.05	0.07	0.04	
Student moved in current school year	0.03	0.03	0.02	
Student moved in previous school year	0.05	0.05	0.00	
Missing student school info for previous year	0.10	0.02	-0.21	
Average classroom peer reading score	0.03	-0.02	-0.06	
Student EOG math score, 5th grade, standardized	0.08	0.05	-0.01	
Student EOG reading score, 5th grade, standardized	0.05	-0.01	-0.03	
Student was limited English proficient in previous year	0.02	0.02	-0.01	

Violent acts per 1,000	2.02	1.21	-0.19
Percentage Asian students	2.46	1.68	-0.17
% AYP goals met	88.51	82.45	-0.14
Percentage Black students	26.75	21.36	-0.13
School is in a mid-sized city	0.16	0.22	0.08
Per-pupil spending on community services	70.89	62.75	-0.02
Per pupil spending on extracurricular activity	23.59	10.64	-0.21
Per pupil spending on food services	482.32	500.98	0.08
Pct of teachers who are fully licensed	0.97	0.98	0.07
Teacher is full time or receives full-time pay	0.54	0.52	-0.02
School ABC status is high performing	0.13	0.12	-0.02
School classified as high growth	0.68	0.72	0.05
Percentage of Hispanic students	10.78	12.15	0.06
School ABC status is low performing	0.31	0.34	0.03
School classified as no growth	0.06	0.05	-0.03
Avg composite EOG passrate (N-passed/N-taken)	0.53	0.53	-0.01
Per pupil spending on professional development	68.22	66.43	-0.02
Avg proportion of teachers with advanced degrees	0.28	0.28	-0.01
Per pupil spending on school leadership	506.65	518.31	0.03
Avg proportion of National Board Certified teachers	0.11	0.10	-0.07
Per-pupil spending on student services	386.60	416.34	0.13
School receives Title I funds	0.66	0.63	-0.04
Total per pupil expenditures	84.50	88.07	0.13
Per pupil spending on transportation	283.18	292.75	0.04
Avg proportion of teachers that returned the following year	0.80	0.84	0.20
Difference 4th-5th grades, % AYP goals met	-3.89	-6.96	-0.09
Difference 4th-5th grades, % Hispanic students	0.57	1.22	0.16
Difference 4th-5th grades, Per-	-61.27	-65.67	-0.06

pupil spending on miscellaneous				
Difference 4th-5th grades, school avg composite EOG passrate	-0.14	-0.13	0.03	
Difference 4th-5th grades, per-pupil spending on special instruction	104.99	77.56	-0.06	
Violent acts per 1,000	10.29	9.72	-0.04	
Percentage Asian students	2.27	1.73	-0.15	
Percentage of AYP goals met	86.91	83.67	-0.16	
Avg. proportion of teachers with 3 or fewer years exp	0.24	0.23	-0.04	
Per pupil spending on food services	430.48	439.06	0.04	
Pct of teachers who are fully licensed	0.91	0.92	0.08	
Per-pupil spending on miscellaneous	59.06	73.26	0.19	
School avg composite EOG passrate	0.64	0.62	-0.06	
Per pupil spending on professional development	68.15	78.54	0.16	
Avg proportion of teachers with advanced degrees	0.27	0.26	-0.03	
Avg proportion of teachers with supplemental Masters degrees	0.19	0.20	0.06	
Per pupil spending on school leadership	520.89	517.90	-0.01	
Avg proportion of National Board Certified teachers	0.08	0.06	-0.27	*
Short-term suspension rate	33.26	31.31	-0.06	
Avg proportion of teachers that return to the school in following year	0.75	0.76	0.08	

Table 3.11.

Balance after PSA Model (Variables Not in PSA Model—Supplement)

	Inverse Probability Weighted Means			
	Comparison	**Treatment**	**Difference (d)**	**d > |.25|**
N per group	81217	6184		
Student is American Indian	0.01	0.00	-0.03	
Overage students in classroom at 75th percentile	0.00	0.00	0.02	
Avg standardized peer EOG score	0.87	0.89	0.08	
Birthdate below standard cutoff for cohort	0.01	0.00	-0.03	
Per-pupil spending on district administration	240.58	371.03	0.46	*
School is located in a large city	0.11	0.00	-0.34	*
Classroom standard deviation of previous year EOG math scores	0.87	0.88	0.03	
School is a magnet	0.07	0.07	0.01	
Per pupil spending on miscellaneous	0.08	-0.01	-0.09	
School ABC status is No Recognition	0.02	0.01	-0.01	
Avg proportion of teachers that returned from previous year	0.87	0.92	0.33	*
School is located in a rural area	0.50	0.46	-0.05	
Avg proportion of teachers with supplemental Masters	0.20	0.21	0.07	
Years of teacher experience	11.51	11.71	0.01	
School has a year-round academic calendar	0.08	0.00	-0.28	*
Difference 4th-5th grades, violent acts per 1,000	0.27	0.48	0.04	
Difference 4th-5th grades, % of students who are Asian	0.06	-0.04	-0.06	
Difference 4th-5th grades, % of students who are Black	-0.15	-0.70	-0.11	
Difference 4th-5th grades, per-pupil spending on capital outlays	-168.26	-573.99	-0.09	
Difference 4th-5th grades, per-pupil spending on community services	21.28	47.43	0.09	
Difference 4th-5th grades, per	27.88	44.84	0.14	

pupil spending on food services			
Difference 4th-5th grades, % of students receiving free/reduced price lunc	0.22	1.10	0.05
Difference 4th-5th grades, proportion of teachers who returned from previo	-0.01	0.04	0.21
Difference 4th-5th grades, avg proportion of teachers with advanced degree	0.01	0.02	0.12
Difference 4th-5th grades, avg proportion of teachers with supplemental Ma	0.00	0.01	0.09
Difference 4th-5th grades, per pupil spending on school leadership	25.66	39.22	0.10
Difference 4th-5th grades, Avg proportion of National Board Certfified tea	0.01	0.01	0.03
Difference 4th-5th grades, short-term suspension rate	0.13	0.37	0.02
Difference 4th-5th grades, per-pupil spending on student services	-54.72	-30.79	0.07
Difference 4th-5th grades, per pupil spending on transportation	30.78	24.95	-0.11
Difference 4th-5th grades, avg proportion of teachers who return the follo	0.03	0.01	-0.10
Per pupil spending on capital outlay	358.50	88.64	-0.17
Per pupil spending on community services	6.67	2.36	-0.16
Per pupil spending on extracurricular activities	30.58	45.04	0.20
Avg proportion of teachers that returned to the school from previous year	0.88	0.87	-0.01

Table 3.12.

Sample Description after Weighting

	Mean	SD	Min	Median	Max	N(Miss)
Student EOG math score, 3rd grade, difference from 5th	-0.01	0.62	-4.23	-0.01	3.91	0
Student EOG reading score, 3rd grade, difference from 5th	-0.02	0.65	-3.93	-0.02	3.95	0
Student EOG math score, 4th grade, difference from 5th	0.01	0.57	-3.23	0.00	3.11	0
Student EOG reading score, 4th grade, difference from 5th	-0.01	0.61	-3.48	-0.01	3.41	0
Classroom in 75th percentile academically/intellectually gifted	0.02	0.13	0.00	0.00	1.00	0
Student is Asian	0.02	0.14	0.00	0.00	1.00	0
Student is Black	0.27	0.44	0.00	0.00	1.00	0
No of days attended	164.57	15.41	1.00	166.00	186.00	0
Classroom in 75th percentile learning disabled	0.00	0.07	0.00	0.00	1.00	0
Classroom in 75th percentile of students receiving free or reduced price l	0.09	0.29	0.00	0.00	1.00	0
Student receives free or reduced price lunch	0.40	0.49	0.00	0.00	1.00	0
Student is Hispanic	0.09	0.29	0.00	0.00	1.00	0
Student has limited English proficiency	0.05	0.23	0.00	0.00	1.00	0
Student moved in current school year	0.02	0.15	0.00	0.00	1.00	0
Student moved in previous school year	0.05	0.21	0.00	0.00	1.00	0
Missing student school info for previous year	0.08	0.27	0.00	0.00	1.00	0
Average classroom peer reading score	0.01	0.48	-3.17	0.03	1.76	0
Student EOG math score, 5th grade, standardized	0.07	0.97	-2.97	0.06	2.57	0
Student EOG reading score, 5th grade, standardized	0.04	0.98	-2.97	0.11	2.66	0
Student was limited English proficient in previous year	0.02	0.14	0.00	0.00	1.00	0
Violent acts per 1,000	1.85	3.04	0.00	0.00	137.93	0
Percentage Asian students	2.32	3.43	0.00	1.26	33.33	0

% AYP goals met	87.89	14.33	0.82	92.30	100.00	0
Percentage Black students	25.96	22.52	0.00	19.10	99.19	0
School is in a mid-sized city	0.18	0.38	0.00	0.00	1.00	0
Per-pupil spending on community services	69.07	122.61	-12.96	2.44	975.64	0
Per pupil spending on extracurricular activity	20.57	43.71	-4.23	6.25	586.22	0
Per pupil spending on food services	486.63	138.50	17.48	477.88	2107.47	0
Pct of teachers who are fully licensed	0.97	0.03	0.71	0.98	1.00	0
Teacher is full time or receives full-time pay	0.54	0.50	0.00	1.00	1.00	0
School ABC status is high performing	0.13	0.33	0.00	0.00	1.00	0
School classified as high growth	0.69	0.46	0.00	1.00	1.00	0
Percentage of Hispanic students	11.20	10.50	0.00	7.35	70.29	0
School ABC status is low performing	0.32	0.47	0.00	0.00	1.00	0
School classified as no growth	0.07	0.25	0.00	0.00	1.00	0
Avg composite EOG passrate (N-passed/N-taken)	0.53	0.15	0.00	0.53	0.93	0
Per pupil spending on professional development	68.16	40.78	11.42	57.60	755.34	0
Avg proportion of teachers with advanced degrees	0.28	0.10	0.00	0.27	0.65	0
Per pupil spending on school leadership	503.40	158.64	65.80	466.59	2426.18	0
Avg proportion of National Board Certified teachers	0.10	0.08	0.00	0.09	0.50	0
Per-pupil spending on student services	391.69	135.87	48.02	370.02	1853.75	0
School receives Title I funds	0.66	0.47	0.00	1.00	1.00	0
Total per pupil expenditures	85.10	13.57	59.53	82.75	177.71	0
Per pupil spending on transportation	285.51	118.77	17.20	271.84	1892.68	0
Avg proportion of teachers that returned the following year	0.81	0.11	0.00	0.83	1.00	0
Difference 4th-5th grades, % AYP goals met	-3.74	11.72	-47.60	0.00	46.20	0
Difference 4th-5th grades, % Hispanic students	0.72	2.20	-10.48	0.52	11.72	0
Difference 4th-5th grades, Per-pupil spending on miscellaneous	-62.46	36.15	-309.07	-51.91	0.00	0
Difference 4th-5th grades, school avg composite EOG	-0.14	0.06	-0.47	-0.14	0.18	0

passrate						
Difference 4th-5th grades, per-pupil spending on special instruction	98.73	266.51	-7666.32	102.81	2267.54	0
Violent acts per 1,000	10.21	7.74	0.00	8.14	47.62	0
Percentage Asian students	2.18	2.66	0.00	1.33	18.70	0
Percentage of AYP goals met	86.18	10.66	52.00	88.00	100.00	0
Avg. proportion of teachers with 3 or fewer years exp	0.24	0.10	0.00	0.23	0.55	0
Per pupil spending on food services	432.63	131.92	-0.21	427.04	1131.40	0
Pct of teachers who are fully licensed	0.91	0.07	0.53	0.93	1.00	0
Per-pupil spending on miscellaneous	61.65	36.88	14.75	47.23	237.43	0
School avg composite EOG passrate	0.63	0.13	0.18	0.64	1.00	0
Per pupil spending on professional development	70.50	33.75	20.85	58.81	257.64	0
Avg proportion of teachers with advanced degrees	0.27	0.09	0.04	0.27	0.52	0
Avg proportion of teachers with supplemental Masters degrees	0.20	0.07	0.02	0.19	0.42	0
Per pupil spending on school leadership	519.83	142.31	239.69	493.76	1523.45	0
Avg proportion of National Board Certified teachers	0.08	0.06	0.00	0.06	0.35	0
Short-term suspension rate	33.18	22.49	0.62	27.14	199.66	0
Avg proportion of teachers that return to the school in following year	0.75	0.10	0.43	0.76	1.00	0
Student is American Indian	0.01	0.09	0.00	0.00	1.00	0
Overage students in classroom at 75th percentile	0.00	0.05	0.00	0.00	1.00	0
Avg standardized peer EOG score	0.87	0.16	0.08	0.89	1.93	0
Birthdate below standard cutoff for cohort	0.01	0.08	0.00	0.00	1.00	0
Per-pupil spending on district administration	268.02	126.97	90.62	223.48	1194.23	0
School is located in a large city	0.09	0.28	0.00	0.00	1.00	0
Classroom standard deviation of previous year EOG math scores	0.87	0.18	0.00	0.87	2.07	0
School is a magnet	0.07	0.25	0.00	0.00	1.00	93

Per pupil spending on miscellaneous	0.07	0.91	-5.08	0.00	11.38	0
School ABC status is No Recognition	0.02	0.12	0.00	0.00	1.00	0
Avg proportion of teachers that returned from previous year	0.89	0.09	0.25	0.90	1.00	0
School is located in a rural area	0.49	0.50	0.00	0.00	1.00	0
Avg proportion of teachers with supplemental Masters	0.20	0.08	0.00	0.19	0.59	0
Years of teacher experience	11.55	9.68	0.00	9.00	49.00	5776
School has a year-round academic calendar	0.06	0.25	0.00	0.00	1.00	0
Difference 4th-5th grades, violent acts per 1,000	0.38	3.27	-20.00	0.00	69.44	0
Difference 4th-5th grades, % of students who are Asian	0.03	0.91	-6.08	0.00	5.31	0
Difference 4th-5th grades, % of students who are Black	-0.21	3.59	-52.24	-0.09	26.12	0
Difference 4th-5th grades, per-pupil spending on capital outlays	-217.78	2293.27	-25754.72	0.00	18694.10	0
Difference 4th-5th grades, per-pupil spending on community services	27.69	116.90	-479.37	0.00	869.76	0
Difference 4th-5th grades, per pupil spending on food services	33.59	81.90	-809.75	42.09	1090.30	0
Difference 4th-5th grades, % of students receiving free/reduced price lunc	0.34	7.25	-92.36	0.60	38.90	0
Difference 4th-5th grades, proportion of teachers who returned from previo	0.00	0.14	-0.59	0.00	0.95	0
Difference 4th-5th grades, avg proportion of teachers with advanced degree	0.01	0.05	-0.19	0.01	0.30	0
Difference 4th-5th grades, avg proportion of teachers with supplemental Ma	0.00	0.05	-0.19	0.00	0.30	0
Difference 4th-5th grades, per pupil spending on school leadership	29.66	83.88	-554.38	24.79	884.18	0
Difference 4th-5th grades, Avg proportion of National Board Certfified tea	0.01	0.04	-0.22	0.00	0.22	0
Difference 4th-5th grades, short-term suspension rate	0.16	5.87	-168.46	0.05	78.65	0
Difference 4th-5th grades, per-pupil spending on student	-45.93	117.08	-695.76	-28.14	625.24	0

services						
Difference 4th-5th grades, per pupil spending on transportation	30.17	35.32	-1653.42	26.18	226.08	0
Difference 4th-5th grades, avg proportion of teachers who return the follo	0.02	0.13	-0.86	0.01	0.75	493
Per pupil spending on capital outlay	304.50	1505.90	-12.51	30.88	19989.74	0
Per pupil spending on community services	5.82	16.71	-0.03	0.00	130.77	0
Per pupil spending on extracurricular activities	33.79	35.74	0.00	29.76	283.10	0
Avg proportion of teachers that returned to the school from previous year	0.88	0.12	0.00	0.89	1.00	0

Table 3.13A.

Treatment Effects (Questions)

		Q1: Overall CS (PSA) Impact (Sig)	Q2: Comparison of PSA Impact with RCT Impact					Q3D: RCT control - nonRCT	Q3S: Effects of RCT Control and nonRCT	
			RCT Impact	Difference	Test	prob > |t|		Diff (Sig)	RCTControl (Sig)	NonRCT (Sig)
Math										
	Main	-0.04	0.039	-0.079	-7.9000	0.000	***	0.03	-0.02	-0.05
R1	Not Imputed	-0.04	0.039	-0.079	-7.9000	0.000	***	0.04	-0.02	-0.06
R2	Group 4	-0.03	0.039	-0.069	-6.9000	0.000	***	0.01	-0.02	-0.03
R3	No Group 3	-0.02	0.039	-0.059	-5.9000	0.000	***	0.00	-0.02	-0.02
R4	No Groups 2, 3	-0.02	0.039	-0.059	-5.9000	0.000	***	0.01	-0.02	-0.03
R5	ML	-0.05	0.039	-0.085	-8.5000	0.000	***	0.09	-0.02	0.07
R6	No Movers	-0.05	0.039	-0.085	-8.5000	0.000	***	0.10	-0.02	-0.07
Reading										
	Main	0.04	0.018	0.022	0.9565	0.339		-0.02	0.03	0.05
		(p < .05)								(p < .05)
R1	Not Imputed	0.03	0.018	0.012	0.5217	0.602		0.01	0.03	0.03
R2	Group 4	0.02	0.018	0.002	0.0870	0.931		0.01	0.03	0.02
R3	No Group 3	0.04	0.018	0.022	0.9565	0.339		-0.03	0.03	0.06
		(p < .05)								(p < .05)
R4	No Groups 2, 3	0.03	0.018	0.012	0.5217	0.602		-0.01	0.03	0.04
R5	ML	0.03	0.018	0.012	0.5217	0.602		0.01	0.03	0.03
R6	No Movers	0.03	0.018	0.012	0.5217	0.602		0.01	0.03	0.03

Table 3.13b.

Credibility of the PSA Design

	PSA Design	Difference	Test	prob > |t|	
Math					
Main	0.00	0.039	3.90	0.000	***
R1	0.00	0.039	3.90	0.000	***
R2	0.00	0.039	3.90	0.000	***
R3	0.00	0.039	3.90	0.000	***
R4	0.00	0.039	3.90	0.000	***
R5	0.00	0.039	3.90	0.000	***
R6	0.00	0.039	3.90	0.000	***
Reading					
Main	-0.01	0.028	1.22	0.224	
R1	-0.01	0.028	1.22	0.224	
R2	-0.01	0.028	1.22	0.224	
R3	-0.01	0.028	1.22	0.224	
R4	-0.01	0.028	1.22	0.224	
R5	-0.01	0.028	1.22	0.224	
R6	0.00	0.018	0.78	0.434	

Table 3.14.

Math (Overall PSA Treatment Effect—Imputed Data)

	Unconditional					**Conditional**				
	3 Level Random Slope					**3 Level Random Slope**				
	Est.	**SE**	**t value**	**p value**		**Est.**	**SE**	**t value**	**p value**	
Intercept	0.04	0.02	1.96	0.050		0.07	0.01	6.00	0.00	***
Developmental Time	0.00	0.00	-6.51	0.000	***	0.00	0.00	-5.83	0.00	***
Developmental Time-Squared	0.00	0.00	-4.93	0.000	***	0.00	0.00	-5.55	0.00	***
Interruption/Transition to MS	-0.01	0.01	-2.24	0.025	*	-0.01	0.01	-2.39	0.02	*
Treatment School (PSA Design)	-0.06	0.08	-0.74	0.459		-0.04	0.05	-0.71	0.48	
DD: Treatment (PS) by Interruption	-0.04	0.03	-1.53	0.126		-0.04	0.03	-1.52	0.13	
Classroom in 75th percentile academically/intellectually gifted						-0.11	0.02	-4.46	0.00	***
Student is Black						-0.36	0.01	-52.72	0.00	***
No of days attended						0.00	0.00	7.86	0.00	***
Classroom in 75th percentile learning disabled						0.03	0.04	0.78	0.44	
Student is academically/intellectually gifted						0.99	0.01	148.18	0.00	***
Student has learning disability						-0.43	0.01	-53.37	0.00	***
Classroom in 75th percentile of students receiving free or reduced price l						0.01	0.01	1.07	0.29	
Student receives free or						-0.20	0.01	-35.37	0.00	***

reduced price lunch					
Student is Hispanic	-0.05	0.01	-4.03	0.00	***
Student has limited English proficiency	-0.28	0.01	-20.85	0.00	***
Average classroom peer math score	0.20	0.01	16.16	0.00	***
Student is a male	0.05	0.00	10.55	0.00	***
Student moved in current school year	-0.08	0.02	-5.24	0.00	***
Student moved in previous school year	-0.09	0.01	-8.28	0.00	***
Average classroom peer reading score	0.06	0.01	4.83	0.00	***
Student was limited English proficient in previous year	0.22	0.02	13.23	0.00	***
School ABC status is high performing	0.23	0.04	5.65	0.00	***
School classified as high growth	0.09	0.02	3.84	0.00	***
School ABC status is low performing	-0.14	0.03	-4.41	0.00	***
School classified as no growth	-0.04	0.04	-1.13	0.26	
Overage students in classroom at 75th percentile	-0.12	0.06	-2.14	0.03	*
Birthdate below standard cutoff for cohort	0.15	0.03	5.36	0.00	***
Classroom standard deviation of previous year EOG math scores	0.01	0.02	0.64	0.52	
School is a magnet	0.22	0.04	5.07	0.00	***
Classroom standard deviation	-0.05	0.02	-2.84	0.00	**

of previous year EOG math scores						
Years of teacher experience		0.00	0.00	-0.42	0.68	
Violent acts per 1,000		0.00	0.00	-1.58	0.11	
Average days in attendance		0.00	0.01	-0.18	0.86	
Percentage Black students		0.00	0.00	-4.99	0.00	***
Avg. proportion of teachers with 3 or fewer years exp		0.17	0.13	1.27	0.20	
Pct of teachers who are fully licensed		0.27	0.18	1.49	0.14	
Percentage Hispanic students		0.00	0.00	0.47	0.64	
Percentage free/reduced lunch students at school		-0.01	0.00	-5.66	0.00	***
Per pupil-spending on regular instruction		0.00	0.00	-1.16	0.25	
Avg proportion of teachers that returned to the school from previous year		-0.20	0.12	-1.64	0.10	
Avg proportion of teachers with advanced degrees		0.86	0.24	3.65	0.00	***
Avg proportion of teachers with supplemental Masters degrees		-0.56	0.26	-2.17	0.03	*
Avg proportion of National Board Certified teachers		0.29	0.18	1.62	0.11	
Per-pupil expenditures		0.00	0.00	1.09	0.27	
Short-term suspension rate		0.00	0.00	0.26	0.80	
School receives Title I funds		0.03	0.03	0.93	0.35	
Residual	0.11	0.11				
School random intercept	0.12	0.04				
School intercept-slope	-0.01	-0.01				

covariance		
School random slope	0.01	0.01
Student random intercept	0.72	0.39
Student intercept-slope covariance	-0.08	-0.08
Student random slope	0.15	0.15

Table 3.15.

Math (Difference in PSA Treatment Effect between RCT Control and Non-RCT—Imputed Data)

	Unconditional					**Conditional**				
	3 Level Random Slope					**3 Level Random Slope**				
	Est.	**SE**	**t value**	**p value**		**Est.**	**SE**	**t value**	**p value**	
Intercept	0.04	0.02	1.88	0.060		0.07	0.01	5.88	0.00	***
RCT District School	-0.06	0.17	-0.32	0.753		0.07	0.10	0.65	0.51	
Developmental Time	0.00	0.00	-6.72	0.000	***	0.00	0.00	-5.95	0.00	***
Developmental Time-Squared	0.00	0.00	-5.10	0.000	***	0.00	0.00	-5.56	0.00	***
Interruption/Transition to MS	-0.01	0.01	-2.20	0.028	*	-0.01	0.01	-2.31	0.02	*
Treatment School (PSA Design)	-0.04	0.11	-0.37	0.713		-0.06	0.07	-0.93	0.35	
DD: Treatment (PS) by Interruption	-0.05	0.03	-1.54	0.123		-0.05	0.03	-1.53	0.13	
DDD: RCT District School by Interruption	0.03	0.05	0.61	0.543		0.03	0.05	0.61	0.54	
Classroom in 75th percentile academically/intellectually gifted						-0.11	0.02	-4.38	0.00	***
Student is Black						-0.36	0.01	-53.03	0.00	***
No of days attended						0.00	0.00	7.71	0.00	***
Classroom in 75th percentile learning disabled						0.05	0.04	1.26	0.21	
Student is academically/intellectually gifted						0.99	0.01	148.67	0.00	***
Student has learning						-0.44	0.01	-53.72	0.00	***

disability					
Classroom in 75th percentile of students receiving free or reduced price l	0.02	0.01	1.41	0.16	
Student receives free or reduced price lunch	-0.20	0.01	-35.49	0.00	***
Student is Hispanic	-0.05	0.01	-4.10	0.00	***
Student has limited English proficiency	-0.28	0.01	-20.87	0.00	***
Average classroom peer math score	0.20	0.01	16.05	0.00	***
Student is a male	0.05	0.00	10.57	0.00	***
Student moved in current school year	-0.08	0.02	-5.39	0.00	***
Student moved in previous school year	-0.09	0.01	-8.23	0.00	***
Average classroom peer reading score	0.07	0.01	5.43	0.00	***
Student was limited English proficient in previous year	0.22	0.02	13.29	0.00	***
School ABC status is high performing	0.23	0.04	5.76	0.00	***
School classified as high growth	0.09	0.02	4.04	0.00	***
School ABC status is low performing	-0.15	0.03	-4.47	0.00	***
School classified as no growth	-0.05	0.04	-1.25	0.21	
Overage students in classroom at 75th percentile	-0.12	0.05	-2.33	0.02	*
Birthdate below standard cutoff for cohort	0.15	0.03	5.57	0.00	***

Classroom standard deviation of previous year EOG math scores		-0.01	0.02	-0.92	0.36	
School is a magnet		0.22	0.04	5.14	0.00	***
Years of teacher experience		0.00	0.00	-0.44	0.66	
Violent acts per 1,000		0.00	0.00	-1.72	0.09	
Average days in attendance		0.00	0.01	-0.30	0.77	
Percentage Black students		0.00	0.00	-5.06	0.00	***
Avg. proportion of teachers with 3 or fewer years exp		0.17	0.13	1.32	0.19	
Pct of teachers who are fully licensed		0.26	0.18	1.45	0.15	
Percentage Hispanic students		0.00	0.00	0.48	0.63	
Percentage free/reduced lunch students at school		-0.01	0.00	-5.59	0.00	***
Per pupil-spending on regular instruction		0.00	0.00	-1.30	0.19	
Avg proportion of teachers that returned to the school from previous year		-0.21	0.12	-1.75	0.08	
Avg proportion of teachers with advanced degrees		0.86	0.24	3.68	0.00	***
Avg proportion of teachers with supplemental Masters degrees		-0.59	0.26	-2.26	0.02	*
Avg proportion of National Board Certified teachers		0.30	0.18	1.67	0.09	
Per-pupil expenditures		0.00	0.00	1.19	0.24	
Short-term suspension rate		0.00	0.00	0.27	0.79	
School receives Title I funds		0.03	0.03	0.90	0.37	
Residual	0.11	0.11				

School random intercept	0.13	0.04
School intercept-slope covariance	-0.01	-0.01
School random slope	0.01	0.01
Student random intercept	0.72	0.39
Student intercept-slope covariance	-0.08	-0.08
Student random slope	0.15	0.15

Table 3.16.

Math (Treatment Effects for RCT Control, Non-RCT and RCT Treatment—Imputed Data)

	Unconditional					Conditional				
	3 Level Random Slope					3 Level Random Slope				
	Est.	**SE**	**t value**	**p value**		**Est.**	**SE**	**t value**	**p value**	
Intercept	0.04	0.02	1.88	0.061		0.07	0.01	5.44	0.00	***
Developmental Time	0.00	0.00	-8.15	0.000	***	0.00	0.00	-7.07	0.00	***
Developmental Time-Squared	0.00	0.00	-5.56	0.000	***	0.00	0.00	-5.81	0.00	***
Interruption/Transition to MS	-0.01	0.01	-1.98	0.047	*	-0.01	0.01	-2.10	0.04	*
DD: Non-RCT School by Interruption	-0.05	0.03	-1.49	0.135		-0.05	0.03	-1.52	0.13	
DD: RCT Control School by Interruption	-0.02	0.04	-0.47	0.632		-0.02	0.04	-0.46	0.65	
DD: RCT Treatment School by Interruption	0.00	0.04	0.02	1.000		0.00	0.04	0.10	0.91	
Non-RCT School	-0.04	0.11	-0.37	0.711		-0.06	0.07	-0.85	0.39	
RCT Control School	-0.10	0.14	-0.70	0.484		0.01	0.09	0.11	0.91	
RCT Treatment School	-0.02	0.14	-0.15	0.880		-0.02	0.09	-0.28	0.78	
Classroom in 75th percentile academically/intellectually gifted						-0.10	0.02	-4.28	0.00	***
Student is Black						-0.37	0.01	-53.06	0.00	***
No of days attended						0.00	0.00	7.71	0.00	***
Classroom in 75th percentile learning						0.06	0.04	1.49	0.14	

disabled					
Student is academically/intellectually gifted	0.97	0.01	145.65	0.00	***
Student has learning disability	-0.44	0.01	-53.62	0.00	***
Classroom in 75th percentile of students receiving free or reduced price l	0.01	0.01	1.00	0.32	
Student receives free or reduced price lunch	-0.19	0.01	-34.92	0.00	***
Student is Hispanic	-0.05	0.01	-4.12	0.00	***
Student has limited English proficiency	-0.29	0.01	-21.39	0.00	***
Average classroom peer math score	0.21	0.01	16.34	0.00	***
Student is a male	0.05	0.00	11.08	0.00	***
Student moved in current school year	-0.09	0.01	-5.81	0.00	***
Student moved in previous school year	-0.09	0.01	-8.36	0.00	***
Average classroom peer reading score	0.07	0.01	4.95	0.00	***
Student was limited English proficient in previous year	0.22	0.02	13.23	0.00	***
School ABC status is high performing	0.24	0.04	5.51	0.00	***
School classified as high growth	0.09	0.02	3.92	0.00	***
School ABC status is low	-0.16	0.03	-4.52	0.00	***

performing					
School classified as no growth	-0.04	0.04	-1.10	0.27	
Overage students in classroom at 75th percentile	-0.15	0.05	-2.88	0.00	**
Birthdate below standard cutoff for cohort	0.14	0.03	5.21	0.00	***
Classroom standard deviation of previous year EOG math scores	0.02	0.02	0.92	0.36	
School is a magnet	0.18	0.04	4.14	0.00	***
Classroom standard deviation of previous year EOG math scores	-0.05	0.02	-3.02	0.00	**
Years of teacher experience	0.00	0.00	-0.45	0.65	
Violent acts per 1,000	0.00	0.00	-1.14	0.26	
Average days in attendance	0.00	0.01	-0.57	0.57	
Percentage Black students	0.00	0.00	-4.55	0.00	***
Avg. proportion of teachers with 3 or fewer years exp	0.17	0.14	1.20	0.23	
Pct of teachers who are fully licensed	0.20	0.19	1.08	0.28	
Percentage Hispanic students	0.00	0.00	0.40	0.69	
Percentage free/reduced lunch students at school	-0.01	0.00	-5.52	0.00	***
Per pupil-spending on regular instruction	0.00	0.00	-1.55	0.12	

Avg proportion of teachers that returned to the school from previous year		-0.21	0.13	-1.66	0.10	
Avg proportion of teachers with advanced degrees		0.86	0.25	3.43	0.00	***
Avg proportion of teachers with supplemental Masters degrees		-0.57	0.28	-2.06	0.04	*
Avg proportion of National Board Certified teachers		0.35	0.19	1.84	0.07	
Per-pupil expenditures		0.00	0.00	1.20	0.23	
Short-term suspension rate		0.00	0.00	0.49	0.62	
School receives Title I funds		0.02	0.03	0.82	0.41	
Residual	0.09	0.10				
School random intercept	0.13	0.05				
School intercept-slope covariance	-0.01	-0.01				
School random slope	0.01	0.01				
Student random intercept	0.74	0.40				
Student intercept-slope covariance	-0.07	-0.07				
Student random slope	0.15	0.15				

Table 3.17.

Reading (Overall PSA Treatment Effect—Imputed Data)

	Unconditional 3 Level Random Slope					**Conditional 3 Level Random Slope**				
	Est.	**SE**	**t value**	**p value**		**Est.**	**SE**	**t value**	**p value**	
Intercept	0.02	0.02	1.48	0.140		0.04	0.01	5.87	0.00	***
Developmental Time	-0.01	0.00	-14.99	0.000	***	-0.01	0.00	-15.12	0.00	***
Developmental Time-Squared	0.00	0.00	-1.18	0.237		0.00	0.00	-2.26	0.02	*
Interruption/Transition to MS	0.00	0.00	0.63	0.530		0.00	0.00	1.09	0.28	
Treatment School (PSA Design)	-0.13	0.07	-1.87	0.062		-0.11	0.03	-3.25	0.00	**
DD: Treatment (PS) by Interruption	0.04	0.02	2.43	0.015	*	0.04	0.02	2.31	0.02	*
Classroom in 75th percentile academically/intellectually gifted						-0.11	0.03	-4.17	0.00	***
Student is Black						-0.34	0.01	-47.28	0.00	***
No of days attended						0.00	0.00	5.19	0.00	***
Classroom in 75th percentile learning disabled						0.07	0.04	1.55	0.12	
Student is academically/intellectually gifted						0.86	0.01	123.44	0.00	***
Student has learning disability						-0.54	0.01	-63.09	0.00	***

Classroom in 75th percentile of students receiving free or reduced price l	-0.02	0.01	-1.76	0.08	
Student receives free or reduced price lunch	-0.23	0.01	-39.04	0.00	***
Student is Hispanic	-0.10	0.01	-8.22	0.00	***
Student has limited English proficiency	-0.45	0.01	-31.62	0.00	***
Average classroom peer math score	0.10	0.01	7.37	0.00	***
Student is a male	-0.08	0.01	-14.78	0.00	***
Student moved in current school year	-0.09	0.02	-5.98	0.00	***
Student moved in previous school year	-0.05	0.01	-4.71	0.00	***
Average classroom peer reading score	0.17	0.01	12.23	0.00	***
Student was limited English proficient in previous year	0.19	0.02	10.90	0.00	***
School ABC status is high performing	0.16	0.03	5.52	0.00	***
School classified as high growth	0.00	0.02	-0.26	0.80	
School ABC status is low performing	-0.11	0.02	-4.57	0.00	***
School classified as no growth	0.00	0.03	-0.16	0.87	
Overage students in classroom at 75th percentile	-0.21	0.06	-3.58	0.00	***

Birthdate below standard cutoff for cohort	0.12	0.03	4.20	0.00	***
Classroom standard deviation of previous year EOG math scores	0.04	0.02	2.17	0.03	*
School is a magnet	0.20	0.03	6.62	0.00	***
Classroom standard deviation of previous year EOG math scores	-0.07	0.02	-4.09	0.00	***
Years of teacher experience	0.00	0.00	0.76	0.45	
Violent acts per 1,000	0.00	0.00	-1.12	0.26	
Average days in attendance	0.00	0.00	-0.65	0.52	
Percentage Black students	0.00	0.00	-5.04	0.00	***
Avg. proportion of teachers with 3 or fewer years exp	-0.02	0.10	-0.21	0.83	
Pct of teachers who are fully licensed	0.17	0.13	1.32	0.19	
Percentage Hispanic students	0.00	0.00	-2.79	0.01	**
Percentage free/reduced lunch students at school	-0.01	0.00	-8.51	0.00	***
Per pupil-spending on regular instruction	0.00	0.00	-1.21	0.23	
Avg proportion of teachers that returned to the school from previous year	-0.06	0.08	-0.75	0.45	
Avg proportion of teachers with advanced degrees	0.64	0.17	3.85	0.00	***
Avg proportion of teachers	-0.52	0.18	-2.84	0.00	**

with supplemental Masters degrees					
Avg proportion of National Board Certified teachers		0.15	0.13	1.15	0.25
Per-pupil expenditures		0.00	0.00	1.35	0.18
Short-term suspension rate		0.00	0.00	-1.17	0.24
School receives Title I funds		0.00	0.02	0.18	0.86
Residual	0.12	0.11			
School random intercept	0.08	0.01			
School intercept-slope covariance	0.01	0.00			
School random slope	0.00	0.00			
Student random intercept	0.73	0.42			
Student intercept-slope covariance	-0.08	-0.07			
Student random slope	0.15	0.15			

Table 3.18.

Reading (Difference in PSA Treatment Effect between RCT Control and Non-RCT— Imputed Data)

	Unconditional					**Conditional**				
	3 Level Random Slope					**3 Level Random Slope**				
	Est.	**SE**	**t value**	**p value**		**Est.**	**SE**	**t value**	**p value**	
Intercept	0.02	0.02	1.43	0.152		0.04	0.01	5.76	0.00	***
RCT District School	-0.07	0.14	-0.46	1.000		-0.01	0.07	-0.14	0.89	
Developmental Time	-0.01	0.00	-15.31	0.000	***	-0.01	0.00	-15.23	0.00	***
Developmental Time-Squared	0.00	0.00	-1.23	0.220		0.00	0.00	-2.36	0.02	*
Interruption/Transition to MS	0.00	0.00	0.63	0.531		0.00	0.00	1.06	0.29	
Treatment School (PSA Design)	-0.10	0.09	-1.16	0.248		-0.11	0.04	-2.46	0.01	*
DD: Treatment (PS) by Interruption	0.05	0.02	2.23	0.026	*	0.05	0.02	2.13	0.04	*
DDD: RCT District School by Interruption	-0.02	0.04	-0.64	0.525		-0.02	0.03	-0.52	0.60	
Classroom in 75th percentile academically/intellectually gifted						-0.10	0.03	-3.99	0.00	***
Student is Black						-0.34	0.01	-47.39	0.00	***
No of days attended						0.00	0.00	5.06	0.00	***
Classroom in 75th percentile learning disabled						0.08	0.04	2.00	0.05	*
Student is academically/intellectually						0.86	0.01	123.43	0.00	***

gifted					
Student has learning disability	-0.54	0.01	-63.27	0.00	***
Classroom in 75th percentile of students receiving free or reduced price l	-0.02	0.01	-1.37	0.17	
Student receives free or reduced price lunch	-0.23	0.01	-39.11	0.00	***
Student is Hispanic	-0.10	0.01	-8.30	0.00	***
Student has limited English proficiency	-0.45	0.01	-31.59	0.00	***
Average classroom peer math score	0.09	0.01	6.99	0.00	***
Student is a male	-0.08	0.01	-14.80	0.00	***
Student moved in current school year	-0.09	0.02	-6.14	0.00	***
Student moved in previous school year	-0.05	0.01	-4.83	0.00	***
Average classroom peer reading score	0.18	0.01	13.27	0.00	***
Student was limited English proficient in previous year	0.19	0.02	10.96	0.00	***
School ABC status is high performing	0.16	0.03	5.44	0.00	***
School classified as high growth	0.00	0.02	-0.30	0.76	
School ABC status is low performing	-0.10	0.02	-4.43	0.00	***
School classified as no growth	0.00	0.03	-0.16	0.87	

Overage students in classroom at 75th percentile	-0.21	0.06	-3.69	0.00	***
Birthdate below standard cutoff for cohort	0.13	0.03	4.40	0.00	***
Classroom standard deviation of previous year EOG math scores	0.00	0.02	0.05	0.96	
School is a magnet	0.19	0.03	6.44	0.00	***
Years of teacher experience	0.00	0.00	0.66	0.51	
Violent acts per 1,000	0.00	0.00	-1.09	0.28	
Average days in attendance	0.00	0.00	-0.64	0.52	
Percentage Black students	0.00	0.00	-4.91	0.00	***
Avg. proportion of teachers with 3 or fewer years exp	-0.02	0.10	-0.26	0.80	
Pct of teachers who are fully licensed	0.17	0.13	1.32	0.19	
Percentage Hispanic students	0.00	0.00	-2.75	0.01	**
Percentage free/reduced lunch students at school	-0.01	0.00	-8.35	0.00	***
Per pupil-spending on regular instruction	0.00	0.00	-1.19	0.23	
Avg proportion of teachers that returned to the school from previous year	-0.06	0.08	-0.74	0.46	
Avg proportion of teachers with advanced degrees	0.64	0.17	3.77	0.00	***
Avg proportion of teachers	-0.52	0.19	-2.78	0.01	**

with supplemental Masters degrees					
Avg proportion of National Board Certified teachers		0.14	0.13	1.11	0.27
Per-pupil expenditures		0.00	0.00	1.33	0.18
Short-term suspension rate		0.00	0.00	-1.14	0.26
School receives Title I funds		0.00	0.02	0.15	0.88
Residual	0.11	0.11			
School random intercept	0.08	0.01			
School intercept-slope covariance	0.01	0.00			
School random slope	0.00	0.00			
Student random intercept	0.73	0.42			
Student intercept-slope covariance	-0.08	-0.07			
Student random slope	0.15	0.15			

Table 3.19.

Reading (Treatment Effects for RCT Nontrol, Non-RCT and RCT Treatment— Imputed Data)

	Unconditional					**Conditional**				
	3 Level Random Slope					**3 Level Random Slope**				
	Est.	**SE**	**t value**	**p value**		**Est.**	**SE**	**t value**	**p value**	
Intercept	0.02	0.02	1.48	0.138		0.04	0.01	4.91	0.00	***
Developmental Time	-0.01	0.00	-14.95	0.000	***	-0.01	0.00	-14.45	0.00	***
Developmental Time-Squared	0.00	0.00	-2.67	0.008	**	0.00	0.00	-3.55	0.00	***
Interruption/Transition to MS	0.00	0.00	0.23	0.822		0.00	0.00	0.57	0.57	
DD: Non-RCT School by Interruption	0.05	0.02	2.20	0.028	*	0.05	0.02	2.06	0.04	*
DD: RCT Control School by Interruption	0.03	0.03	1.05	0.296		0.03	0.03	1.10	0.27	
DD: RCT Treatment School by Interruption	-0.01	0.03	-0.39	0.693		-0.01	0.03	-0.38	0.71	
Non-RCT School	-0.10	0.09	-1.14	0.255		-0.11	0.05	-2.05	0.04	*
RCT Control School	-0.17	0.11	-1.51	0.130		-0.11	0.06	-1.74	0.08	
RCT Treatment School	-0.08	0.11	-0.75	0.438		-0.05	0.06	-0.76	0.45	
Classroom in 75th percentile academically/intellectually gifted						-0.09	0.02	-3.84	0.00	***
Student is Black						-0.34	0.01	-48.17	0.00	***
No of days attended						0.00	0.00	5.20	0.00	***
Classroom in 75th percentile learning disabled						0.10	0.04	2.37	0.02	*

Student is academically/intellectually gifted	0.85	0.01	122.86	0.00	***
Student has learning disability	-0.54	0.01	-63.95	0.00	***
Classroom in 75th percentile of students receiving free or reduced price l	-0.02	0.01	-1.92	0.06	
Student receives free or reduced price lunch	-0.22	0.01	-38.69	0.00	***
Student is Hispanic	-0.10	0.01	-8.51	0.00	***
Student has limited English proficiency	-0.46	0.01	-32.52	0.00	***
Average classroom peer math score	0.10	0.01	7.58	0.00	***
Student is a male	-0.08	0.01	-14.94	0.00	***
Student moved in current school year	-0.11	0.02	-7.42	0.00	***
Student moved in previous school year	-0.03	0.01	-3.22	0.00	**
Average classroom peer reading score	0.17	0.01	12.60	0.00	***
Student was limited English proficient in previous year	0.19	0.02	11.00	0.00	***
School ABC status is high performing	0.15	0.04	4.33	0.00	***
School classified as high growth	0.01	0.02	0.30	0.77	
School ABC status is low performing	-0.12	0.03	-4.28	0.00	***

School classified as no growth	0.01	0.03	0.16	0.88	
Overage students in classroom at 75th percentile	-0.26	0.06	-4.62	0.00	***
Birthdate below standard cutoff for cohort	0.12	0.03	4.20	0.00	***
Classroom standard deviation of previous year EOG math scores	0.04	0.02	2.37	0.02	*
School is a magnet	0.16	0.04	4.40	0.00	***
Classroom standard deviation of previous year EOG math scores	-0.07	0.02	-4.22	0.00	***
Years of teacher experience	0.00	0.00	0.78	0.44	
Violent acts per 1,000	0.00	0.00	-0.41	0.68	
Average days in attendance	0.00	0.00	-0.97	0.33	
Percentage Black students	0.00	0.00	-3.74	0.00	***
Avg. proportion of teachers with 3 or fewer years exp	-0.03	0.12	-0.22	0.82	
Pct of teachers who are fully licensed	0.10	0.16	0.66	0.51	
Percentage Hispanic students	0.00	0.00	-2.00	0.05	*
Percentage free/reduced lunch students at school	-0.01	0.00	-7.37	0.00	***
Per pupil-spending on regular instruction	0.00	0.00	-1.36	0.18	
Avg proportion of teachers	-0.06	0.10	-0.56	0.58	

that returned to the school from previous year						
Avg proportion of teachers with advanced degrees		0.63	0.21	3.06	0.00	**
Avg proportion of teachers with supplemental Masters degrees		-0.48	0.23	-2.12	0.03	*
Avg proportion of National Board Certified teachers		0.20	0.16	1.28	0.20	
Per-pupil expenditures		0.00	0.00	1.19	0.24	
Short-term suspension rate		0.00	0.00	-0.51	0.61	
School receives Title I funds		0.00	0.02	0.16	0.87	
Residual	0.11	0.10				
School random intercept	0.08	0.02				
School intercept-slope covariance	0.01	0.00				
School random slope	0.00	0.00				
Student random intercept	0.74	0.42				
Student intercept-slope covariance	-0.07	-0.07				
Student random slope	0.16	0.16				

Table 3.20.

Intensity of Implementation in RCT and PSM Evaluation Periods: Number of CareerStart Lessons Used by School and Subject

Math	**RCT**			**PSM**	
	6	**7**	**8**	**6**	**7**
350	30	34	26	27	24
392	24	17	6	19	20
396	11	13	6	7	4
406	24	31	24	57	27
464	23	26	29	21	34
480	3	4	1	10	17
492	20	11	10	10	10

Language Arts	**RCT**			**PSM**	
	6	**7**	**8**	**6**	**7**
350	31	27	27	10	26
392	22	13	12	23	17
396	13	15	3	17	11
406	19	24	27	45	33
464	17	21	33	28	30
480	3	8	14	7	17
492	11	8	0	9	16

Language Arts	**RCT**			**PSM**	
	6	**7**	**8**	**6**	**7**
350	29	16	22	22	21
392	17	17	15	23	21
396	14	13	4	4	2
406	17	16	21	41	25
464	24	29	35	26	30
480	9	5	6	7	16
492	9	12	1	5	6

Language Arts	**RCT**			**PSM**	
	6	**7**	**8**	**6**	**7**
350	26	17	20	17	22
392	22	24	28	17	25
396	11	12	1	7	5
406	15	19	16	42	25

464	23	27	31	21	32
480	8	7	8	4	18
492	7	6	0	4	4

Seven RCT Treatment Schools
RCT from 2006-07 (6th grade) through 2008-09 (8th grade)
PSM from 2008-09 (6th grade) through 2009-10 (7th grade) (2010-11 not available)

APPENDIX 1. ESTIMATION OF COMPOSITE IN URM

First, I calculated the difference between each student score at time w, w-1 and w-2, and the mean of school means at each time point. Second, I estimated the partitioned covariance matrix C:

(7) $$C = \begin{bmatrix} c_{yy} & c_{yx} \\ c_{xy} & C_{xx} \end{bmatrix}$$

The expectation-maximization algorithm was used to estimate C in the presence of conditionally random missing values. Third, I estimated the coefficients of a projection equation, b, as follows:

(8) $$b = C_{xx}^{-1} c_{xy}$$

Fourth, I estimated the following prediction equation, using the elements of b as the $\hat{\beta}$ which predicts a composite of students' previous test scores, spanning two years and two subjects, that have been recalibrated as pooled-within-teacher fixed effect values (m = math and r = reading).

(9) $$C_i = \hat{\mu}_y + \hat{\beta}_{m1}(x_{im1} - \hat{\mu}_{m1}) + \hat{\beta}_{m2}(x_{im2} - \hat{\mu}_{m2}) + \hat{\beta}_{r1}(x_{ir1} - \hat{\mu}_{r1}) + \hat{\beta}_{r2}(x_{ir2} - \hat{\mu}_{r2})$$

Finally, I substituted the composite C_i into the two-level (students nested in teachers) model.

APPENDIX 2. CAREERSTART EIGHTH GRADE MATHEMATICS LESSON

Percent of Change: What's the Deal?
Developed by Debbie Brooks, Peggy Dickey, and Jan Sullivan

Essential Question: How can the ability to calculate percent of change be important in retail careers?

Learning Outcomes
• Students will practice calculating percent of change.

Teacher planning
Materials Needed
• Worksheet: "What's the Deal?" (Includes answer key.)
• Overhead copy of worksheet
• Calculator (One for each student)
Time required for lesson: One class period

Activities
1) Ask students when they were last in a store that was having a sale. What signs did they see?
(Students may mention seeing signs that said things like "20% off."
2) Using the career information below, discuss careers in retail stores with the class.
3) Discuss what determines whether you have a percent increase or a percent decrease in price.
(The original is either higher or lower than the new amount.)
4) Tell the students that they are going to be in charge of generating a price sheet that lists mark-ups and sales prices for items that are in a retail store.
5) Place students in groups of no more than three for this activity. Hand out "What's the Deal?" worksheets and have students work through the information given. (40 minutes)
6) Use the overhead copy of the worksheet to have each group write answers to share with the class. (5 minutes)
7) Discuss with students what other careers might involve finding percent of change. (5 minutes) If students need help brainstorming, you may choose to access the following websites:
a) The Bureau of Labor Statistics' Occupational Outlook Handbook
b) The Bureau of Labor Statistics' K-12 career home page
c) CareerOneStop.org
8) Career information
a) Some of the information below comes from the Bureau of Labor Statistics' Occupational Outlook Handbook.

Retail Managers
Retail managers supervise salespersons and other store personnel. They are responsible for hiring and training retail employees and oversee the daily operations of a retail store. One of

the daily operations is setting prices for the items or the inventory in the store. When an item is purchased, the store must mark-up the item.
The mark-up needs to be high enough to bring in income, but low enough so the item will sell. When business becomes slow or items do not sell, stores run sales and discount the sales price. Sales managers must be accurate with the mark-ups and discounts.
Retail stores must have a profit to stay in business. The formula used to determine profit is: Profit = Sales - Expenses. Sales are the money customers pay for items purchased. From these sales, expenses are deducted. Expenses include the cost for items or inventory, rent or mortgages, utilities, salaries for salespersons, salaries for office workers, salaries for custodians, advertising costs, office and cleaning supplies, insurance, payroll taxes, sales tax, shipping costs, and any other expense needed to operate the store. The profit will only come from what is left of the mark-up.
• Education: High school and college-level courses related to business
• Pay: $27,500 - $46,500
• Growth: Slower than average; 4% increase over the next 10 years

Salespersons
Salespersons are responsible for assisting customers in finding items to purchase, ringing up the sales, and bagging purchases. They should always be courteous and helpful to the customer. Salespersons may be responsible for opening and closing the cash register and are expected to maintain accurate accounting of the contents in the cash register.
• Education: No formal education is required
• Pay: $17,000 - $41,500
• Growth: Average growth; 12% increase over the next 10 years

REFERENCES

Agodino, R., & Dynarski, M. (2004). Are experiments the only option? A look at dropout prevention programs. *Review of Economics and Statistics*, *86*(1), 180–194.

Akos, P., Konold, T., & Niles, S. (2004). A career readiness typology and typal membership in middle school. *Career Development Quarterly*, *53*(1), 53–66.

Akos, P. & Galassi, J. (2004). Middle and high school transitions as viewed by students, parents, and teachers. *Professional School Counseling*, *7*, 212–221.

Albers, C. A., & Kratochwill, T. R. (2006). Teacher and principal consultations: Best practices. In Franklin, C., Harris, M. B., & Allen-Meares, P. *The School Services Sourcebook: A Guide for School-Based Professionals*. Oxford, UK: Oxford University Press.

Allen, J. M. (2009). Valuing practice over theory: How beginning teachers re-orient their practice in the transition from the university to the workplace. *Teaching and Teacher Education*, *25*, 647–654.

Allison, P. D. (2009). *Fixed Effects Regression Models. Series: Quantitative Applications in the Social Sciences.* Thousand Oaks, CA: Sage.

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology*, *84*(3), 261–271.

Amrein-Beardsley, A., & Collins, C. (2012). The SAS Education Value-Added Assessment System (SAS® EVAAS®) in the Houston Independent School District (HISD): Intended and unintended consequences. *Education Policy Analysis Archives*, *20*(12), 1–29. Retrieved from http://epaa.asu.edu/ojs/article/view/1096/982.

Amrein-Beardsley, A. (2008). Methodological concerns about the Education Value-Added Assessment System. *Educational Researcher*, *37*(2), 75.

Anderman, E. M., Anderman, L. H., Yough, M. S., & Gimbert, B. G. (2010). Value-added models of assessment: Implications for motivation and accountability. *Educational Psychologist*, *45*(2), 123–137.

Anderman, E. M., & Midgley, C. (1997). Changes in personal achievement goals and the perceived goal structures across the transition to middle schools. *Contemporary Educational Psychology*, *22*, 269–298.

Anderman, E. M., Maehr, M. L., & Midgley, C. (1999). Declining motivation after the transition to middle school: Schools can make a difference. *Journal of Research and Development in Education*, *32*, 131–147.

Andrejko, L. (2004). Value-added assessment: A view from a practitioner. *Journal of Educational and Behavioral Statistics*, *29*(1), 7–9.

Angrist, J. D., Imbens, G. W., & Rubin, D. R. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–455. doi:10.2307/2291629

Arellano, M., & Bond, S. (1991) Some tests of specification of panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies*, *58*, 277–298.

Assor, A., Kaplan, H., & Roth, G. (2002). Choice is good, but relevance is excellent: Autonomy-enhancing and suppressing teacher behaviours predicting students' engagement in schoolwork. *British Journal of Educational Psychology*, *72*(2), 261–278.

Balducci, D. E., & Wandner., S. A. (2009). Managing applied research in federal employment policy. *Journal of Policy Analysis and Management*, *28*(1), 166–167.

Balfanz,, R., Herzog, L., & Mac Iver, D. J. (2007). Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. *Educational Psychologist*, *42*(4), 223–235.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*. *29*(1), 37–65.

Beller, E., & Hout, M. (2006). Intergenerational social mobility: The United States in comparative perspective. *Future of Children*, *16*(2), 19–36.

Bifulco, R. (2012). Can nonexperimental estimates replicate estimates based on random assignment in evaluations of school choice? A within-study comparison. *Journal of Policy Analysis and Management*, *31*(3), 729–751.

Birman, B. F., Desimone, L., Porter, A. C., & Garet, M. S. (2001). Designing professional development that works. *Educational Leadership*, *38*(4), 28–33.

Bishop, R., Berryman, M., Wearmouth, J., Peter, M., & Clapham, S. (2012). Professional development, changes in teacher practice and improvement in Indigenous students' educational performance: A case study from New Zealand. *Teaching and Teacher Education*, *28*, 694–705.

Bloom, H. W. (2003). *Sample Design for an Evaluation of the Reading First Program*. New York: MDRC.

Bloom, H. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review*, *8*(2), 225–246.

Borman, K. M., Eitle, T. M., Michael, D., Eitle, D. J., Lee, R., Johnson, L., Cobb-Roberts, D., Dorn, S., & Shircliffe, B. (2004). Accountability in a postdesegregation era: The continuing significance of racial segregation in Florida's schools. *American Educational Research Journal*, *41*(3), 605–631.

Bowen, N. K., Bowen, G. L., & Ware, W. B. (2002). Neighborhood Social Disorganization, Families, and the Educational Behavior of Adolescents. *Journal of Adolescent Research*, *17*, 468–490.

Bowen, G. L., Rose, R. A., & Bowen, N. A. (2005). *The Reliability and Validity of the School Success Profile*. Philadelphia: Xlibris Press.

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., Wyckoff, J. (2008). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis*, *31*(4), 416–440.

Britt, T. W. (2005). The effects of identity-relevance and task difficulty on task motivation, stress, and performance. *Motivation and Emotion*, *29*(3), 189–202.

Bronfenbrenner, U. (1989). Ecological systems theory. *Annals of Child Development*, *6*, 187–249.

Brophy, J. & Good, T. (1986). Teacher behavior and student achievement. In M. Wittrock (Ed.), *Handbook of Research on Teaching* (pp. 3280375). NY: Macmillan.

Burke, M., & Sass, T. (2008). Classroom Peer Effects and Student Achievement. Federal Reserve Bank of Boston Working Paper Series No. 08-5. Boston, MA.

Buzhardt, J., Greenwood, C. R., Abbott, M., & Tapia, Y. (2006). Research on scaling up evidence-based instructional practice: Developing a sensitive measure of the rate of implementation. *Educational Technology Research and Development*, *54*(5), 467–492.

Cameron, C. E., Connor, C. M., Morrison, F. J., & Jewkes, A. M. (2008). Effects of classroom organization on letter-word reading in first grade. *Journal of School Psychology*, *46*, 173–192.

Carlson, D., Borman, G. D., Robins, M. (2011). A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis*, *33*(3), 378–398.

Carnoy, M., Jacobsen, R., Mishel, L., & Rothstein, R. (2006). Worth the price? Weighing the evidence on charter school achievement. *Education Finance and Policy*, *1*(1), 151–161.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). The long-term impact of teachers: Teacher value-added and student outcomes in adulthood. NBER Working Paper Series #17699.

Clarke, J., & Dede, C. (2009). Design for scalability: A case study of the river city curriculum. *Journal of Science Education and Technology*, *18*, 353–365.

Clotfelder, C.T., Glennie, E., Ladd, H. F., & Vigdor, J. L. (2008). Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina. *Journal of Public Economics*, *92*, 1352–1370.

Clotfelder, C. T., Glennie, E., Ladd, H. F., & Vigdor, J.L. (2008). Teacher bonuses and teacher retention in low-performing schools: Evidence from the North Carolina $1,800 teacher bonus program. *Public Finance Review*, *36*(1), 63–87.

Clotfelder, C. T., Ladd, H. F., & Vigdor, J. L. (2010). Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects. *Journal of Human Resources*. *45*(3), 655–681.

Clotfelder, C. T., Ladd, H. F., & Vigdor, J. L. (2005). "Who teachers whom? Race and the distribution of novice teachers. *Economics of Education Review*, *24*, 377–392.

Clotfelder, C. T., Ladd, H. F., Vigdor, J. L., & Diaz, R. A. (2004). Do school accountability systems make it more difficult for low-performing schools to attract and retain high-quality teachers? *Journal of Policy Analysis and Management*, *23*(2), 251–271.

Coburn, C. E. (2003). Rethinking scale: Moving beyond numbers to deep and lasting change. *Educational Researcher*, *32*(6), 3–12.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, *25*(2), 119–142.

Coleman, J., Campbell, E., Hobsen, C., McPartland, J., Mood, A., Weinfeld, F., & York, R. (1966). *Equality of Educational Opportunity*. Washington, DC: U.S. Government Printing Office.

Constas, M. A. (2007). Reshaping the methodological identity of education research: Early signs of the impact of federal policy. *Eval. Rev*., *31*, 391–400

Cook, P. J., MacCoun, R., Muschkin, C., & Vigdor, J. (2008). The negative impacts of starting middle school in sixth grade. *Journal of Policy Analysis and Management*. *27*(1), 104–121.

Cook, T. D., Shadish, W. R., & Wong, V. W. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, *27*(4), 724–750.

Cook, T. D., & Steiner, P. M. (2009). Some empirically viable alternatives to random assignment. *Journal of Policy Analysis and Management*, *28*(1), 165–166.

Cook, T. D. (1999). Considering the major arguments against random assignment: An analysis of the intellectual culture surrounding evaluation in American schools of education. Paper presented at the Harvard Faculty Seminar on Experiments in Education.

Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, *88*(4), 715–730.

Crumpton, H. E. & Gregory, A. (2011). "I'm Not Learning": The role of academic relevancy for low-achieving students. *Journal of Educational Research*, *104*(1), 42–53.

Darling-Hammond, L., Berry, B., & Thoreson, A. (2001). Does teacher certification matter? Evaluating the evidence. *Educational Evaluation and Policy Analysis*, *23*(1),57–77.

Davies, P. (1999). What is evidence-based education? *British Journal of Educational Studies*, *47*(2), 108–121.

Davis, P., & Boruch, R. (2001). The Campbell Collaboration: Does for public policy what Cochrane does for health. *Editorial. British Medical Journal*, *323*, 294–295.

Dehejia, R., and Wahba, R. (1999). Causal effects in non experimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94*(448), 1053–1062.

Destin, M., & Oyserman, D. (2010). Incentivizing education: Seeing schoolwork as an investment, not a chore. *Journal of Experimental Social Psychology*, *46*, 846–849.

Dewey, J. (1916). *Democracy and Education*. Toronto, Ontario: Collier-Macmillan Canada.

Diaz, J. J., & Handa, S. (2006). An assessment of propensity score matching as a nonexperimental impact estimator: Evidence from Mexico's PROGRESA Program. *Journal of Human Resources*, *41*(2), 319–345.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values and goals. *Annual Review of Psychology*, *53*, 109–132.

Elias, M. J., Zins, J. E., Graczyk, P A., & Weissberg, R. P. (2003). Implementation, sustainability, and scaling up of social-emotional and academic innovations in public schools. *School Psychology Review*, *32*(3), 303–319.

Elliott, J. (2001). Making evidence based practice educational. *British Education Research Journal*, *27*(5), 555–574.

Finn, J. D., & Rock, D. A. (1997). Academic success among students at-risk for school failure. *Journal of Applied Psychology*, *82*, 221–234.

Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Education Research Journal*, *27*(3), 557–577.

Finn, J. D. (1989). Withdrawing from school. *Review of Educational Research*, *59*(2), 117–142.

Finnigan, K. S., & Gross, B. (2007). Do accountability policy sanctions influence teacher motivation? Lessons from Chicago's low-performing schools. *American Educational Research Journal*, *44*(3), 594–629.

Fixsen, D. L., Blasé, K. A., Naoom, S. F., & Wallace, F. (2009). Core implementation components. *Research on Social Work Practice*, *19*(5), 531–540.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., et al. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science*, *6*(3), 151–175.

Fleming, N. (2011). Middle schoolers getting prepped for higher education. *Education Week*, Dec 07 2011. Retrieved from http://www.edweek.org/ew/articles/2011/12/07/13middle_ep-2.h31.html.

Frangakis, C. E., & Rubin, D. R. (2002). Principal stratification in causal inference. *Biometrics*, *58*(1), 21–29.

Fraser, M. W., Richman, J. M., Galinsky, M. J., & Day, S. H. (2009). *Intervention Research: Developing Social Programs*. New York: Oxford University Press.

Fredericks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*, 59–109.

Frolich, M. (2004). Finite-sample properties of propensity-score matching and weighting estimators. *Review of Economics and Statistics*, *86*(1), 77–90.

Fuchs, D., & Fuchs, L. S. (1998). Researchers and teachers working together to adapt instruction for diverse learners. *Learning Disabilities Research & Practice*, *13*(3), 126–137.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, *38*(4), 915–945.

Geier, R., Blumenthal, P. C., Marx, R. W., Krajcik, J. S., Fishman, B., Soloway, E., & Clay-Chambers, J. (2008). Standardized test outcomes for students engaged in inquiry-based science curricula in the context of urban reform. *Journal of Research in Science Teaching*, *45*, 922–939.

Gennetian, L. A., Magnuson, K., & Morris, P. A. (2008). From statistical associations to causation: What developmentalists can learn from instrumental variable techniques coupled with experimental data. *Developmental Psychology*, *44*(2), 381–394.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulleting*, *117*(1), 21–38.

Glazerman, S., Mayer, D., & Decker, P. (2006). Alternative routes to teaching: The impacts of Teach for America on student achievement and other outcomes. *Journal of Policy Analysis and Management*, *25*(1), 75–96.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *Annals of the American Academy of Political and Social Sciences, 589, Misleading Evidence and Evidence-Led Policy: Making Social Science More Experimental*, 63–93.

Goldhaber, D., & Chaplin, D. (2012). Assessing the "Rothstein Test": Does it really show teacher value-added models are biased? National Center for Analysis of Longitudinal Data in Education Research Working paper Series.

Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement, *Educational Evaluation and Policy Analysis*, *22*(2), 129–145.

Gordon, R., Kane, T. J., and Staiger, D. O. (2006) Identifying effective teachers using performance on the job. Hamilton Project Discussion Paper. Washington DC: Brookings Institution.

Greene, B. A., Miller, R. B., Corwson, H. M., Duke, B. L., & Akey, K. L. (2004). Predicting high school students' cognitive engagement and achievement: Contributions of classroom perceptions and motivation. *Contemporary Educational Psychology*, *29*(4), 462–482.

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of Educational Research*, *66*(3), 411–416.

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). Interpreting research on school resources and student achievement: A rejoinder to Hanushek. *Review of Educational Research*, *66*(3), 361–396.

Grossman, P. L., Valencia, S. W., Evans, K., Thompson, C., Martin, S., & Place, N. (2000). Transitions into teaching: Learning to teach writing in teacher education and beyond. *Journal of Literacy Research*, *32*(4), 631–662.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2010. Can value-added measures of teacher performance be trusted? Unpublished draft manuscript.

Guo, S., & Hussey, D. (1999). Analyzing longitudinal rating data: A three-level hierarchical linear model. *Social Work Research, 23*(4), 258–269.

Hagay, G., & Baram-Tsabari, A. (2011). A shadow curriculum: Incorporating students' interests into the formal biology curriculum. *Research in Science Education, 41*(5), 611–634.

Hamre, B. K., Pianta, R. C., Burchinal, M., Field, S., LoCasale-Crouch, J., Downer, J. T., Howes, C., LaParo, K., & Scott-Little, C., (2012). A course on effective teacher-child interactions: Effects on teacher beliefs, knowledge, and observed practice. *American Educational Research Journal, 49*(1), 88–123.

Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of Applied Econometrics, 18*, 527–544.

Hanushek, E. A. (1981). Throwing money at schools. *Journal of Policy Analysis and Management, 1*, 19–41.

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature, 24*, 1141–1177.

Harackiewicz, J. M., Barron, K. E., Pintrich, P. R., Elliot, A. J., & Thrash, T. M. (2002). Revision of achievement goal theory: Necessary and illuminating. *Journal of Educational Psychology, 94*(3), 638–645.

Harackiewicz, J. M., & Elliot, A. J. Achievement goals and intrinsic motivation. *Journal of Personality and Social Psychology, 65*(5), 904–915.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Bulletin, 15*(3), 234–249.

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., and Sass, T. R.. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics, 95*(7-8): 798–812.

Harris, D., & Sass, T.R. (2006). Value-added models and the measurement of teacher quality. Unpublished manuscript.

Heckman, J. J. (2008). Role of income and family influence on child outcomes. *Annals of New York Academy of Sciences. 1136*, 307–323.

Heckman, J. J. (1997). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources, 32*, 441–462.

Heckman, J. J., & Smith, J. A. (1995). Assessing the case for social experiments. *Journal of Economic Perspectives*, *9*(2), 85–110.

Hedges, L V., & Hedberg, E. C. (n.d.) Intraclass correlation values for planning group randomized trials in education. Unpublished draft manuscript.

Hedges, L. V., Laine, R. D., & Greenwald, R. (1994). Exchange: Part I: Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Educational Researcher*, *23*(5), 5–24.

Henry, G. T., Rose, R. A., & Lauen, D. L. (Under review). Are value-added models good enough for teacher evaluations? Empirical tests of nine commonly used value-added models with actual and simulated data. *Educational Researcher*.

Henry, G. T., Kershaw, D. C., Zulli, R. A., & Smith, A. A. (2011). Incorporating teacher effectiveness into teacher preparation program evaluation. *Journal of Teacher Education*, *63*(5), 335–355.

Henry, G. T., Bastian, K. C., & Fortner, C. K. (2011). Stayers and leavers: Early-career teacher effectiveness and attrition. *Educational Researcher*, *40*(6), 271–280.

Henry, G. T. & Rickman, D. K. (2007). Do peers influence children's skill development in preschool? *Economics of Education Review*, *26*, 100–112.

Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically unmovitated: A critical issue for the 21st century. *Review of Educational Research*, *70*(2), 151–179.

Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist*, *41*(2), 111–127.

Hiebert, J., & Morris, A. K. (2012). Teaching, rather than teachers, as a path toward improved classroom instruction. *Journal of Teacher Education*, *63*(2), 92–102.

Hill, H. C. (2009). Evaluating value-added models: A validity argument approach. *Journal of Policy Analysis and Management*, *28*(4), 700–712.

Hill, H. C., Charalambous, C., Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56–64.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*(3), 794–831.

Hill, H. C., Rowan, B.., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*(2), 371–406.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Pehlps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*(4), 430–511.

Hill, J., Weiss. C., & Zhai, F. (2011). Challenges with propensity score strategies in a high-dimensional setting and a potential alternative. *Multivariate Behavioral Research*, *46*(3), 477–513

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161–1189.

Hleboswith, P. (2012). When best practices aren't: A Schabian perspective on teaching. *Journal of Curriculum Studies*, *44*(1), 1–12.

Holland, P. (2010). A framework and history for score linking. In N. Dorans, M. Pommerich, and P. Holland (eds.), *Linking and Aligning Scores and Scales* (pp. 5–30). New York: Springer.

Holland, Paul W. (1986) Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–960. doi:10.2307/2289064

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.

Horng, E. L. (2009). Teacher tradeoffs: disentangling teachers' preferences for working conditions and student demographics. *American Educational Research Journal*, *46*(3), 690–717.

Hoyt, K. B. (2005). *Career Education: History and Future*. Tulsa, OK: National Career Development Association.

Hughes, J. N., & Chen, Q. (2011). Reciprocal effects of student-teacher and student-peer relatedness: Effects on academic self-efficacy. *Journal of Applied Developmental Psychology*, *32*, 278–287.

Hughes, J. N., Luo, W., Kwok, O. M., & Loyd, L. K. (2008). Teacher-student support, effortful engagement, and achievement: A 3-year longitudinal study. *Journal of Educational Psychology*, *100*(1), 1–14.

Hulleman, C. S., Durik, A. M., Schweigert, S. A. & Harckiewicz, J. M. (2008). Task values, achievement goals, and interest: An integrative analysis. *Journal of Educational Psychology*, *100*(2), 398–416.

Hulleman, C. S., Godes, O., Hendricks, B. L., & Harackewicz, J. M. (2010). Enhancing interest and performance with a utility value intervention. *Journal of Educational Psychology*, *102*(4), 880–895.

Husman, J., Derryberry, W. P., Crowson, H. M., & Lomax, R. (2004). Instrumentality, task value and intrinsic motivation: Making sense of their independent interdependence. *Contemporary Educational Psychology*, *29*(1), 63–76.

Husman, J., & Lens, W. (1999). The role of the future in student motivation. *Educational Psychologist*, *34*(2), 113–125.

Imbens, G. W., & Lemieux, T. (2007). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*, 615–635.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5–86.

Jacob, B. A. (2011). Do principals fire the worst teachers? *Educational Evaluation and Policy Analysis*, *33*(4), 403–434.

Jacob, B. A., Lefgren, L., & Sims, D. P. (2010). The persistence of teacher-induced learning. *Journal of Human Resources*, *45*(4), 915–943.

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

Jang, H., Kim, E. J., & Reeve, J. (2012). Longitudinal test of self-determination theory's motivation mediation model in a naturally occurring classroom context. *Journal of Educational Pscyhology*, *104*(4), 1175–1188.

Janosz, M., Archambault,I., Morizot, J., & Pagani, L. S. (2008). School engagement trajectories and their differential predictive relations to dropout. *Journal of Social Issues*, *64*, 21–40.

Kagan, D. M. (1992). Professional growth among preservice and beginning teachers. *Review of Educational Research*, *62*(2), 129–169.

Kahle, J. B., Meece, J., & Scantlebury, K. (2000). Urban African-American middle school science students: Does standards-based teaching make a difference? *Journal of Research in Science Teaching*, *37*(9), 1019–1041.

Kang, J. D. Y., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*(4), 523–539.

Kelcey, B. (2011). Covariate selection in propensity scores using outcome proxies. *Multivariate Behavioral Research*, *46*(3), 453–476.

Kelchtermans, G. (2005). Teachers' emotions in educational reforms: Self-understanding, vulnerable commitment and micropolitical literacy. *Teaching and Teacher Education*, *21*, 995–1006.

Keller, J. M. (1987). Development and use of the ARCS model of instructional design. *Journal of Instructional Development*, *10*(3), 2–10. doi: 10.1007/BF02905780

Kennedy, M. M. (2010). Attribution error and the quest for teacher quality. *Educational Researcher*, *39*(8), 591–598.

Kennedy, M. M. (1999). Approximations to indicators of student outcomes. *Educational Evaluation and Policy Analysis*, *21*(4), 345–363.

Kinderman, T. A. (2007). Effects of naturally existing peer groups on changes in academic engagement in a cohort of sixth graders. *Child Development*, *78*(4), 1186–1203.

Koedel, C., & Betts, J. R. (2009). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy*, *6*(1), 18-42.

Koretz, D. (2008). Limitations in the use of achievement tests. Journal of Human *Resources*, *27*(4), 752–777.

Koretz, D. M. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *Journal of Human Resources*, *37*(4), 752-777.

Koth, C. W., Bradshaw, C. P., & Leaf, P. J. (2008). A multilevel study of predictors of student perceptions of school climate: The effect of classroom-level factors. *Journal of Educational Psychology*, *100*(1), 96–104.

Krantz-Kent, R. (2008). Teachers' work patterns: when, where, and how much do U.S. teachers work? *Monthly Labor Review*, March, 52–59.

Krapp, A. (2005). Basic needs and the development of interest and intrinsic motivational orientations. *Learning and Instruction*, *15*, 391–395.

Labaree, D. F. (1997). Public goods, private goods: The American struggle over educational goals. *American Educational Research Journal*, *34*, 39–81.

Ladd, H. (2011). Teachers' perceptions of their working conditions: How predictive of planned and actual teacher movement? *Educational Evaluation and Policy Analysis*, *33*(2), 235–261.

Ladd, H. (2002). School vouchers: a critical view. *Journal of Economic Perspectives*. *16*. 3–24.

Langbeim, L. (2009). Beyond random assignment for internal validity and beyond social research for random assignment. *Journal of Policy Analysis and Management*, *28*(1), 164–181.

Lee, O., & Krajcik, J. (2012). Large-scale interventions in science education for diverse student groups in varied educational settings. *Journal of Research in Science Teaching*, *49*(3), 271–280.

Lee, V. E., & Smith, J. B. (1999). Social support and achievement for young adolescents in Chicago: The role of school academic press. *American Educational Research Journal*, *36*(4), 907–945.

Lefgren, L. (2004). Educational peer effects and the Chicago public schools. *Journal of Urban Economics*, *56*, 169–191.

Lepper, M. R., & Cordova, D. I. (1992). A desire to be taught: Instructional consequences of intrinsic motivation. *Motivation and Emotion*, *16*, 187-–208.

Lepper, M. R., Corpus, J. H., & Iyengar, S. S. (2005). Intrinsic and extrinsic motivational orientations in the classroom: Age differences and academic correlates. *Educational Psychology*, *97*(2), 184–196.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, *29*(2), 4–16.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*, 1198–1202.

Lockwood, J. R., McCaffrey, Mariano, L. T., & Setodji, C. (2007). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics*, *32*(2), 125–150.

Lynch, S. J., Pyke, C., & Grafton B. H. (2012). A retrospective view of a study of middle school science curriculum materials: Implementation, scale-up, and sustainability in a changing policy environment. *Journal of Research in Science Teaching*, *49*(3), 305–332.

Marks, H. M. (2000). Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. *American Educational Research Journal*, *37*, 153–184.

Markus, H. R., & Nurius, P. (1986). Possible selves. *American Psychologist*, *41*, 954–969. doi:10.1037/0003-066X.41.9.954

Marsh, H. W., Overall, J. U., & Kesler, S. P. Validity of student evaluations of instructional effectiveness: A comparison of faculty self-evaluations and evaluations by their students. *Journal of Educational Psychology*, *71*(2), 149–160.

Mashburn, A. J., Justice, L. M., Downer, J. T., & Pianta, R. C. (2009). Effects on children's language achievement during pre-kindergarten. *Child Development*, *80*(3), 686–702.

Matsumura, L. C., Garnier, H. E., & Resnick, L. B. (2010). Implementing literacy coaching: The role of school social resources. *Educational Evaluation and Policy Analysis, 32*(2), 249–272.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67–101.

McDonald, S.-K., Keesler, V. A., Kauffman, N. J., & Schneider, B. (2006). Scaling-up exemplary interventions. *Educational Researcher, 35*(3), 15–24.

Meece, J. L. (2003). Applying learner-centered principals to middle school education. *Theory into Practice, 42*(2), 109–116.

Miller, R. B., & Brickman, S. J. (2004). A model of future-oriented motivation and self-regulation. *Educational Psychology Review, 16*(1), 9–33.

Morgan, S. L., & Winship, C. (2007) *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* Cambridge, UK: Cambridge University Press.

National Research Council. (1999). *High Stakes: Testing for Tracking, Promotion and Graduation*. Washington, DC: National Research Council.

Nye, B., Konstantopolous, S., & Hedges, L. (2004). How large are teacher effects? *Educational Evaluation & Policy Analysis, 26*(3), 237–257.

Nye, B. A., Hedges, L. V., & Konstantopoulos, S. (1999). The long term effects of small classes: A five year followup of the Tennessee Class Size Experiment. *Educational Evaluation and Policy Analysis, 21*, 127–142.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research, 78*(1), 33–84.

Oakley, A. (2002). Social science and evidence-based everything: the case of education. *Educational Review, 54*(3), 277–286.

Orthner, D. K., Jones-Sanpei, H., Akos, P., Rose, R. A. (2012). Improving middle-school student engagement through career-relevant instruction in the core curriculum. *Journal of Educational Research, 106*(1), 27–38.

Orthner, D. K., Akos, P., Rose, R. A., Jones-Sanpei, H., Mercado, M., & Woolley, M. E. (2010). CareerStart: A middle school student engagement and academic achievement program. *Children & Schools, 32*, 223–234.

Oyserman, D. (2008). Racial-ethnic self-schemas: Multidimensional identify-based motivation. *Journal of Research in Personality, 42*(5), 1186–1198.

Oyserman, D., Terry, K., & Bybee, D. (2002). A possible selves intervention to enhance school involvement. *Journal of Adolescence*, *25*, 313–326.

Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research*, *62*(3), 307–332.

Palardy, G. J., & Rumberger, R. W. (2008). Teacher effectiveness in first grade: The importance of background qualifications, attitudes, and instructional practices for student learning. *Educational Evaluation and Policy Analysis*, *30*(2), 111–140.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference.* Cambridge, UK: Cambridge University Press.

Pearson, M. (2013, January 15). Wins, losses and draws in Chicago school strike. Retrieved January 15, 2013, from http://www.cnn.com/2012/09/19/us/illinois-chicago-teachers-strike/index.html.

Peikes, . D. N., Moreno, L., and Orzol, S. M. (2008). Propensity score matching. *American Statistician*, *62*(3), 222-231.

Penuel, W. R., Fishman, B. J., Cheng B. H., & Sabelli, N. (2011). Organizing research and development at the intersection of learning, implementation, and design. *Educational Researcher*, *40*(7), 331–337.

Phalet, K., Andriessen, I., & Lens, W. (2004). How future goals enhance motivation and learning in multicultural classrooms. *Educational Psychology Review*, *16*(1), 59–89.

Phillippo, K. L., & Rose, R. A. (2012). Treatment Fidelity in the CareerStart Randomized Control Trial: A Mixed Methods Implementation Study. Paper presented at 17th Society for Social Work and Research Annual Meeting, Washington D.C.

Pianta, R. C., Belsky, J., Vandergrift, N., Houts, R., Morrison, F. J. (2008). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal*, *45*(2), 365–397.

Pintrich, P. (2000). Multiple goals, multiple pathways: The role of goal orientation in learning and achievement. *Journal of Educational Psychology*, *92*(3), 544–555.

Pirog, M. (2009). The role of random assignment in social policy research. *Journal of Policy Analysis and Management*, *28*(1), 164–165.

Pirog, M. A., Buffardi, A. L., Chrisinger, C. K., Singh, P., & Briney, J. (2009). Are the alternatives to randomized assignment nearly as good? Statistical corrections to nonrandomized evaluations. *Journal of Policy Analysis and Management*, *28*(1), 164–181.

Pohl, S., Steiner, P. S., Eisermann, J., Soellner, R., & Cook, T. D. (2009). Unbiased causal inference from an observational study: Results of a within-study comparison. *Educational Evaluation and Policy Analysis*, *31*(4), 463–479.

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, *29*(1), 121–129.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual Review of Psychology*, *52*, 501–525.

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

Reckase, M. D. (2004). The real world is more complicated than we would like. *Journal of Educational and Behavioral Statistics*, *29*(1), 117–120.

Reiss, J. (2008). *Error in Economics: Towards a More Evidence-Based Methodology*. London: Routledge.

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

Roeser, R. W., Eccles, J. S., & Sameroff, A. J. (2000). School as context of early adolescents' academic and social-emotional development: A summary of research findings. *Elementary School Journal*, *100*(5), 443–471.

Rose, R. A., & Stone, S. I. (2011). Instrumental variable estimation in social work research: A technique for estimating causal effects in nonrandomized settings. *Journal of the Society for Social Work Research*, *2*(2), 76–88.

Rose, R. A. (2009). *CareerStart: Report on Attrition and Randomization*. Chapel Hill, NC: Jordan Institute for Families.

Rose, R. A., Henry, G. T., & Lauen, D. L. (unpublished manuscript). Comparing value-added models for estimating individual teacher effects on a statewide basis: Simulations and empirical analysis.

Rose, R. A., Woolley, M. E., Orthner, D. K., Akos, P. T., & Jones-Sanpei, H. A. (2012). Increasing teacher use of career-relevant instruction: A randomized control trial of CareerStart. *Educational Evaluation and Policy Analysis*, *34*, 295–312.

Rosenbaum, P. R. (2002). *Observational Studies, 2nd Ed.* Springer Series in Statistics. New York: Springer.

Rosenbaum. P. R., & Rubin, D. R. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.

Rothstein, J. (2010). Teacher quality in educational productionn: tracking, decay and student achievement. *Quarterly Journal of Economics, 125*(1), 175–214.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools. *Teachers College Record, 104*(8), 1525–1567.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*. *29*(1), 103–116.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

Rutkowski, D., Rutkowski, L., & Langfield, G. (2012). Reading economics, thinking education: The relevance—and irrelevance—of economic theory for curriculum research. *Journal of Curriculum Sciences, 44*(2), 165–192.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*, 58–78.

Ryan, A. M., & Patrick, H. (2001). The classroom social environment and changes in adolescents' motivation and engagement during middle school. *American Education Research Journal*, *38*(2), 437–460.

Sackett, D. L, Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., & Richardson,

W. S. (1996) Evidence based medicine: What it is and what it isn't. *British Medical Journal*, *312*, 71–72.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997), "The Tennessee Value-Added Assesment System: A Quantitative, Outcomes-Based Approach to Educational Assesment," in J. Millman (ed.), *Grading Teachers, Grading Schools* (pp. 137–162). Thousand Oaks, CA: Sage.

Santoro, D. A. (2011). Good teaching in difficult times: Demoralization in the pursuit of good work. *American Journal of Education*, *118*(1), 1–23.

Sass, T. (2008). The stability of value-added measures of teacher quality and implications for teacher compensation policy. National Center for Analysis of Longitudinal Data in Education Research.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Boca Raton, FL: Chapman Hall/CRC.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147–177.

Schiefele, U. (1991). Interest, learning and motivation. *Educational Psychologist*, *26*(3&4), 299–323.

Schochet, P. Z. (2011). Do typical RCTs of education interventions have sufficient statistical power for linking impacts on teacher practice and student achievement outcomes? *Journal of Educational and Behavioral Statistics*, *36*(4), 441–471.

Schochet, P. Z., & Chiang, H. S. (2011). Estimation and identification of the complier average causal effect in education RCTs. *Journal of Educational and Behavioral Statistics*, *36*(3), 307–345.

Schochet, P. Z. (2005). *Statistical Power for Random Assignment Evaluations of Education Programs*. Princeton, NJ: Mathematica.

Schoenwald, S. J., & Hoagwood, K. (2001). Effectiveness, transportability, and dissemination of interventions: What matters when? *Psychatric Services*, *52*(9), 1190–1197.

Schunk, D. H., Pintrich, P. R., & Meece, J. M. (2008). *Motivation in Education: Theory, Research & Applications*. Upper Saddle River, NJ: Pearson Education.

Scott, N. (2013, January 10). Seattle Teachers Revolt Against High Stakes Test. Retrieved January 15, 2013, from http://ednotesonline.blogspot.com/2013/01/seattle-teachers-revolt-against-high.html.

Sergiovanni, T. J., Burlingame, M., Coombs, F. S., & Thurston, P. W. (1999). Public values and school policy. In Sergiovanni, T. J., Burlingame, M., Coombs, F. S., & Thurston, P. W., Educational Governance and Administration. Boston, MA: Allyn & Bacon.

Shadish, W. R. (2010). Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings. *Psychological Methods*, *15*(1), 3–17.

Shadish, W. R., Clark, M. H., and Steiner, P. M. (2008). Can nonrandomized experiment yield accurate answers? A randomized experiment comparing random and nonrandom assignments rejoinder. *Journal of the American Statistical Association*, *103*(484), 1353–1356.

Shadish. W. R., & Cook, T. D. (2009). The renaissance of field experimentation in evaluating interventions. *Annual Review of Psychology*, *60*, 607–629.

Shadish, W. R., Cook, T. D. & Campbell. D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. New York: Houghton Mifflin.

Shulman, L. (1987) Knowledge and teaching: foundations of the new reform. *Harvard Educational Review*, *57*, 1–22.

Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the year. *Journal of educational psychology*, *85*(4), 571.

Slavin, R. E. (2008). Perspectives on evidence-based research in education—what works? Issues in synthesizing educational program evaluations. *Educational Researcher*, *37*(1), 5–14.

Smith, T. M., Desimone, L. M., & Ueno, K. (2005). "Highly qualified" to do what? The relationship between NCLB teacher quality mandates and the use of reform-oriented instruction in middle school mathematics. *Educational Evaluation and Policy Analysis*, *27*, 75–109.

Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, *125*, 305–353.

Snijders, T. A. B., and Bosker, R. J. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage.

Springer, M. G., Hamilton, L., McCaffrey, D. F., Ballou, D., Le, V. N., Pepper, M., et al. (2010). Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching. Nashville, TN: National Center on Performance Incentives, Vanderbilt University.

Steiner, P. M., Cook, T. D., & Shadish, W. R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*, *36*(2), 213–236.

Stevenson, J. (2011). Commentary: A contribution to evidence-informed education policy—reflections on Strong, Torgerson, Torgerson and Hulme (2011). *Journal of Chlid Psychology and Psychiatry*, *52*(3), 236–237.

Stipek, D., & Miles, S. (2008). Effects of aggression on achievement: Does conflict with the teacher make it worse? *Child Development*, *79*(6), 1721–1735.

Stuart, E. A., & Green, K. M. (2008). Using full matching to estimate causal effects in nonexperimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology*, *44*(2), 395–406.

Stuart, E. A. (2007). Estimating causal effects using school-level data sets. *Educational Researcher*, *36*(4), 187–198.

Sykes, G., Bird, T., & Kennedy, M. (2010). Teacher education: Its problems and some prospects. *Journal of Teacher Education*, *61*(5), 464–476.

Tatar, D., Roschelle, J., Knudsen, J., Shechtman, N., Kaput, J., & Hopkins, B. (2008). Scaling up innovative technology-based mathematics. *Journal of the Learning Sciences*, *17*, 248–286.

Taylor, E. S., & Tyler, J. H. (2011). The Effect of Evaluation on Performance: Evidence from Longitudinal Student Achievement Data of Mid-Career Teachers. NBER Working Paper Series #16877.

Tekwe, C. D., Carter, R. L., Ma, C. X., Algina, J., Lucas, M. E. et al. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics*, *29*(1), 11–36.

Thoemmes, F. J., & West, S. G. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research*, *46*(3), 514543.

Todd, P. E. & Wolpin, K. I. (2007).The production of cognitive achievement in children: Home, school, and racial test score gaps. *Journal of Human Capital*, *1*(1), 91–136.

Todd, P. E., & Wolpin, K. I. (2003.) On the specification and estimation of the production function for cognitive achievement. *Economic Journal*, *113*, f3–f33.

Torff, B., & Sessions, D. N. (2005). Principals' perceptions of the causes of teacher ineffectiveness. *Journal of Educational Psychology*, *97*(4), 530–537.

Valli, L., & Buese, D. (2007). The changing roles of teachers in an era of high-stakes accountability. *American Educational Research Journal*, *44*(3), 519–558.

Valli, L., Croninger, R. G., & Walters, K. (2007). Who (else) is the teacher? Cautionary notes on teacher accountability systems. *American Journal of Education*, *113*(4), 635–662.

Van Driel, J. H., & Berry, A. (2012). Teacher professional development focusing on pedagogical content knowledge. *Educational Researcher*, *41*(1), 26–28.

Van Ryzen, M. J., Gravely, A. A., & Roseth, C. J. (2009). Autonomy, belongingness, and engagement in school as contributors to adolescent psychological well-being. *Journal of Youth and Adolescence*, *38*, 1–12.

Vignoles, V. L., Golledge, J., Regalia, C., Manzi, C., & Scabini, E. (2006). Beyond self-esteem: Influence of multiple motives on identity construction. *Journal of Personality and Social Psychology*, *90*(2), 308–333.

Voelkl, K. E. (1996). Measuring students' identification with school. *Educational and Psychological Measurement*, *56*(5), 760–770

Wang, J., Spalding, E., Odell, S. J., Klecka, C. L., & Lin, E. (2010). Bold ideas for improving teacher education and teaching: What we see, hear and think. *Journal of Teacher Education*, *61*(1-2), 3–15.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research*, *73*(1), 89–122.

Webster, R. S. (2009). How evidence-based teaching practices are challenged by a Deweyan approach to education. Asia-Pacific Journal of Teacher Education, 37(2), 215-227.

Weiner, B. (1990). History of motivational research in education. *Journal of Educational Psychology*, *82*(4), 616–622.

Wentzel, K. R., & Wigfield, A. (2007). Motivational interventions that work: Themes and remaining issues. *Educational Psychologist*, *42*(4), 261–271.

White, S. P., Wright, J. T., Sanders, W. L., & Rivers, J. C. (2010). SAS EVAAS Statistical Models. Cary, NC: The SAS Institute.

Wigfield, A., & Eccles, J. (1992). The development of achievement task values: A theoretical analysis. *Developmental Review*, *12*, 265–310.

Wilde, E. T., & Hollister, R. (2007). How close is close enough? Evaluating propensity score matching using data from a class size reduction experiment. *Journal of Policy Analysis and Management*, *26*(3), 455–477.

Willett, J. B., Singer, J. D., & Martin, N. C. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations. *Development and Psychopathology*, *10*, 395–426.

Woolley, M. E. & Bowen, G. L. (2007). In the context of risk: Supportive adults and the school engagement of middle school students. *Family Relations*, *56*, 92–104.

Woolley, M. E., Rose, R. A., Mercado, M., & Orthner, D. K. (2013). Teachers teaching differently: A qualitative study of implementation fidelity to professional development. *Journal of Education and Training Studies*, *1*(1), 1–14.

Woolley, M. E. & Rose, R. A. (2012). Measuring teacher instructional practices by student reports: Systematic threats to reliability. Unpublished draft manuscript.

Woolley, M. E., Rose, R. A., Orthner, D., & Jones-Sanpei, H. (In press). Advancing Academic Achievement through Career Relevance in the Middle Grades: A Longitudinal Evaluation of CareerStart. *American Education Research Journal*.

Woolley, M. E., Unick, G. J., Rose, R. A. & Orthner, D. K. (Revise & resubmit). Lasting effects of career relevant instruction in the middle grades: Academic achievement and progress toward graduation in high school. *Instruction and Learning*.

Xu, Z., Hannaway, J., & Taylor, C. (2011). Making a difference? The effects of Teach for America in high school. *Journal of Policy Analysis and Management*, *30*(3), 447–469.

Zimmer, R. W., & Toma, E. F. (2000). Peer effects in private and public schools across countries. *Journal of Policy Analysis and Management*, *19*(1), 75–92.