Title

The Effect of Grading in School Accountability Systems: An Investigation Using Propensity Scores In Second-order Growth Models

Author

Tsui, Jason

Publication Date

2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

The Effect of Grading in School Accountability Systems: An Investigation Using Propensity Scores In Second-order Growth Models

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Education

by

Jason Tsui

2018

ABSTRACT OF THE DISSERTATION

The Effect of Grading in School Accountability Systems:

An Investigation Using Propensity Scores In

Second-order Growth Models

by

Jason Tsui

Doctor of Philosophy in Education

University of California, Los Angeles, 2018

Professor Michael H. Seltzer, Co-chair

Professor Noreen M. Webb, Co-chair

The development and implementation of school-level accountability systems has been mandated by recent federal law. However, there is a dearth of research into the construct validity of such measurement systems. This project adopts a latent factor perspective to assess the validity of a unidimensional definition of School Quality and estimate the impact of implementing a school accountability system using A-F grades for one application: The New York City Progress Report. A novel combination of propensity score matching and second-order latent growth modeling with adjusted error estimates is used. Results show receipt of a failing grade increases School Quality in the second year by 0.167 standard units compared with similar schools. The unidimensional definition of School Quality exhibits extremely poor model fit

however, but model-based grades exhibit better consistency with other external measures of schools compared with the original formulation.

The dissertation of Jason Tsui is approved.

Li Cai

Todd M. Franke

Michael H. Seltzer, Committee Co-chair

Noreen M. Webb, Committee Co-chair

University of California, Los Angeles

2018

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

ATE – Average Treatment Effect

ATT – Average Treatment effect on the Treated

ATU – Average Treatment effect on the Untreated

ELA – English Language Arts

ELL – English Language Learner

ESSA – Every Student Succeeds Act of 2016

FGM – First-order latent Growth Model

NCLB – No Child Left Behind Act of 2001

NYCDOE – New York City Department Of Education

NYCPR – New York City Progress Report

RMSEA – Root Mean Squared Error of Approximation

PS-SGM – Propensity Score adjusted Second-order latent Growth Model

SEM – Structural Equation Modeling

SGM – Second-order latent Growth Model

SUTVA – Stable Unit Treatment Value Assumption

Vita

Education

2006	Master of Arts in Secondary Mathematics Education, Teachers College, Columbia University, New York, NY.
2005	Bachelor of Science in Chemistry, Bachelor of Arts in Physics, University of Chicago, Chicago, IL.

Work

2017 – Present	Data Scientist, Edwire, Inc., El Paso, TX.
2011 – Present	Graduate Student Researcher, University of California, Los Angeles.
2009 – 2011	Mathematics Teacher and Department Head, School of the Future, New York, NY.
2006 – 2009	Mathematics Teacher and Technology Coordinator, Middle School 322, New York City, NY.

Chapter 1. Introduction

Recent trends in educational policy have been characterized by an increased focus on accountability at all levels. Since the passage of The No Child Left Behind of 2001 (NCLB), school-level accountability systems have increasingly been used to pressure schools labeled as failing with mandated interventions, sanctions, or even closure (Hanushek & Raymond, 2005). Several studies have investigated the effect of implementing such systems, both positive and negative (e.g. Carnoy & Loeb, 2002; Dee & Jacob, 2011; Grissmer, Flanagan, Kawata, & Williamson, 2000; Koretz, 2009). But the validity of these measurements remains an issue, making them tenuous bases for decisions at best (e.g. Chay, McEwan, & Urquiola, 2005; Kane & Staiger, 2002). This is especially worrisome as there is evidence that the negative effects of such systems may disproportionately fall on high-poverty schools with diverse student enrollments (Kim & Sunderman, 2005).

Given these findings, the central role of accountability systems in determining pedagogical, financial and personnel outcomes in school is under more and more debate (e.g. Mintrop & Sunderman, 2009). Even so, their development and implementation has been mandated at the state level by the recent update to NCLB, the Every Student Succeeds Act of 2016 (ESSA). Under the current legislation, the bottom 5% of schools identified by each state's system is mandated for a range of interventions (Darling-Hammond et al., 2016).

The purpose of this study is to investigate the efficacy of school interventions based on such measurement systems, especially in light of the diversity of contexts schools may face. Toward this end, I apply statistical techniques to estimate the effect of labeling schools as failing, independent of the influence of demographics. There are three questions guiding this research:

Q1: How accurate are characterizations of school quality?

Q2: How effective is labeling a school as failing at improving school quality?

Q3: Are there demographic differences in school quality?

In order to address these questions, a combination of second-order latent factor growth modeling and propensity score matching provides the statistical framework for the development and comparison of different measurement schema such as might be used in school accountability systems, as well as tools to assess their validity. Data from the New York City Progress Report provide the opportunity to investigate the questions above by adopting this framework.

First implemented in 2006, the New York City Progress Report (NYCPR) is one of the earliest examples of accountability systems that assign grades to every school, based on an array of measurements. Failing grades for schools carry both official consequences—such as school closings and increased scrutiny—and also unofficial consequences, as the results are widely publicized and distributed to parents. Grading systems thus attempt to isolate the targeted sanctions to failing to improve student performance, while also reducing these negative externalities for schools that are doing well.

As a district serving over one million students, the New York City public schools provide a rich and diverse context but also a difficult measurement challenge. Whether this system as specified accurately classifies failing schools is unclear. Very little research has been done into the validity of such systems (Murray & Howe, 2017). Moreover, ascertaining how to answer this question is not straightforward. One difficulty is that there do not exist alternative, objective measures of school quality by which we can assess misclassification rates. There is an analogous

debate occurring around teacher value-added models where, to address such issues, more sophisticated statistical models are being explored (see Glazerman et al., 2010 for a useful discussion on classification errors).

What follows in Chapter 2. is an overview of the trends in educational policy in the United States as pertains to measurement systems for accountability—and in particular, how this movement is rooted in the recognition that the public education system provides inequitable opportunities for students of different backgrounds. Chapter 3 then summarizes the statistical literature pertaining to the proposed methodology.

Chapter 2. A Historical Policy Overview

School evaluation systems in the United States have been growing increasingly common since the late 20th century. In 1983, a report by the National Commission on Excellence entitled *A Nation at Risk: The Imperative for Educational Reform*, painted a dramatic picture of a public education system on the brink of disaster, incapable of neither addressing the changing needs of the nation in a globalizing world nor ameliorating the racial inequality highlighted in the landmark *Brown v. Board of Education* decision of 1954 and the Civil Rights Act of 1964. As one of the recommendations for how to address this crisis, the commission recommended, "[s]tandardized tests of achievement … should be administered at major transition points from one level of schooling to another… The tests should be administered as part of a nationwide (but not Federal) system of State and local standardized tests" (Denning, 1983, p. 125). This was one of the earliest calls for the implementation of standardized testing as a means to provide accountability and remediation to perceived problems within public education.

The prevalence and popularity of standardized tests within accountability systems has only grown since then. Hanushek and Raymond (2005) report that, from 1993 to 2002, the percentage of states with school-level report cards based on standardized test scores grew from less than 10% to nearly 90%. Of these, almost 60% of states had "consequential accountability" systems, defined as systems that attach "consequences such as monetary awards or takeover threats to school performance," reliant almost exclusively on standardized test results. These types of tests became known as "high-stakes," due to the associated consequences. The No Child Left Behind Act (NCLB), signed into law in January of 2002, then enshrined such consequential systems in federal legislation. States now were required to set annual performance goals for each

school; each successive year a school did not "make adequate yearly progress" (AYP) resulted in increasingly invasive interventions. A school that continually failed make AYP faced "restructuring," essentially the equivalent to being shut down. Performance, according to these goals, was defined in terms of "annual measurable objectives" based on proficiency ratings on standardized tests in reading and mathematics. This prioritization of standardized test scores solidified their central role as accountability systems grew in importance and prevalence.

The implementation of NCLB had profound impacts on the educational landscape in the United States. One stated purpose of the legislation was to "ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education" (No Child Left Behind Act of 2001, §§1001). This was accomplished through newly-mandated disaggregation of results by a variety of subgroups, including minority racial and ethnic groups, and students with special educational needs. By highlighting the disparities between different groups of students in stark relief, NCLB forced many educational communities to confront in a public way disparities in educational opportunities and outcomes in their midst.

There were several common criticisms of NCLB, however. The law measured student success by English and math—and eventually science—standardized test scores, as well as high school graduation rates. This set of measures was often decried as too limited, in both breadth and depth, resulting in negative unintended consequences. (For a more in-depth discussion of criticisms of NCLB, see e.g. Darling-Hammond (2007).) A review of the literature suggests some common but unintended consequences of relying primarily on math and English standardized tests, such as a re-distribution of resources and attention away from non-tested subjects such as art (e.g. Klein, Hamilton, McCaffrey, & Stecher, 2000; Ladd & Zelli, 2002; Stecher, Barron, Chun, & Ross, 2000).

Another oft-criticized provision of NCLB required all students to be proficient in mathematics and English by 2014. As results began being reported from districts all over the country, a consensus grew that this expectation was unrealistic. In the interim while Congress worked on an update to NCLB to address such issues, a stop-gap measure called Race To The Top was included in the American Recovery and Reinvestment Act of 2009. This program created a competition for additional funding to individual states in exchange for specific types of policy reforms. At stake were $4.35 billion in additional funding, to be awarded to states based on voluntary applications proposing changes to educational policies. The merit of each application was weighed according to a point system that focused on adoption of common standards and assessments; improvements in performance evaluations for teachers and principals; expansion of charter school opportunities; and interventions for low-performing schools. These reforms sped the development of accountability systems and associated consequences, reinforcing their ascendancy in the educational policy landscape.

The successor and reauthorization to NCLB, The Every Student Succeeds Act (ESSA), was passed in December of 2015, with many of the previous concerns in mind. Of particular relevance to this project, ESSA creates the opportunity and challenge for each state to design its own system for measuring school and student success. The law requires that such systems include multiple measures—alleviating the over-reliance on math and English standardized testing—allowing flexibility in which measures are used and the relative weights with which they are counted. Additionally, for primary and middle schools, the law requires the inclusion of at least one non-traditional measure, for example: student engagement, educator engagement, access to and completion of advanced coursework, post-secondary readiness, or school climate/safety. The law also designates three categories of schools specially mandated for

intervention, the criteria for which require identifying the bottom 5% of schools according to each state's measurement system. The specifics of each system are up to each state to design, which means the problem of how to combine an increasingly diverse selection of measurements is of increasing importance under this new legislative regime. Thus, providing statistical tools to use in the design of such systems meets a timely need, especially to provide feedback regarding their precision and equitability across subgroups. (For a more detailed discussion of the implications of ESSA on school accountability systems and design, see Darling-Hammond et al. (2016).)

2.1. Impact of School Accountability Systems

Beyond the consequences due to legislation, school accountability systems have been shown to have many additional effects—both intended and unintended. Several studies link the use of school accountability systems to increases in student achievement (Carnoy & Loeb, 2002; Dee & Jacob, 2011; Grissmer et al., 2000; Hanushek & Raymond, 2005). While it is difficult to isolate the causal effects of implementing school grading systems on student outcomes, the general consensus is optimistic. Many studies have found statistically significant increases in standardized test scores associated with implementation of school accountability grades (Hanushek & Raymond, 2005; Jacob, 2005; Springer, 2008). Although there is concern that such improvement may be due to undesirable behavior, such as "teaching to the test" (Koretz, 2009) or even blatant cheating (Booher-Jennings, 2005; Jacob & Levitt, 2003), studies have also found corresponding increases in low-stakes test scores where such negative behavior is not incentivized (Dee & Jacob, 2011; Figlio & Rouse, 2006; Grissmer et al., 2000; Jacob, 2005; Klein et al., 2000).

The underlying argument for these grading systems is that by allowing interventions to be targeted to failing schools, students in those schools will be better served. There is evidence to support the claim that there is in fact differential improvement for the lowest-graded schools. Florida implemented an accountability system that provided monetary assistance and training for schools that received an "F" grade. (However, a second "F" within four years resulted in sanctions such as allowing student transfers with vouchers and principal replacement.) Chiang (2009) found that receiving the first "F" was associated with gains in the high-stakes math and reading scores, although not in low-stakes test scores. There was also an increased use of assessments to guide instruction, hiring of subject specialists, implementation of after-school or weekend supplemental instruction, and emphasis on conflict resolution and behavioral interventions. Similarly, by comparing schools on the cusp of receiving failing grades, Rouse, Hannaway, Goldhaber, and Figlio (Rouse, Hannaway, Goldhaber, & Figlio, 2013) estimated receiving an "F" to be associated with at least 15% of test gains in reading and 44% of test gains in math. Rockoff and Turner (2010), in examining the accountability system in New York City, also found that receiving an "F" correlated with a 0.05 to 0.1 standard deviation increase in math and reading scores and a 0.4 standard deviation increase in math and reading score growth.

In contrast, other research has found evidence of unintended negative consequences to accountability systems that may disproportionately affect minority or low-income students. Figlio and Lucas (2004) found that a school receiving high grades tended to attract more wealthy families to its neighborhood, raising housing prices. The implication is that children from economically mobile families had greater access to higher quality education. This would also suggest the converse is true: children from high-poverty families would have more restricted access to higher quality education. Figlio and Kenny (2009) also found that school grades were

correlated with community financial support. Receiving a "D" or "F" dropped contributions by 2/3 or more, with schools serving predominantly poor or minority families especially sensitive to these effects. This reduction in financial support would make it even more difficult for a struggling school to institute reforms. The sanctions associated with failing grades—school closure or student transfers—are also shown to be correlated with negative effects for students. Hanushek, Kain, and Rivkin (2004) found transfers within district—the type caused by sanctions—incur significant short-run academic costs, particularly for poor and minority students. Engberg, Gill, Zamarro, and Zimmer (2012) also found persistent negative effects on academic outcomes for students transferred due to school closings.

These issues all highlight the importance of accurately identifying excellent and low-performing schools in such accountability systems. At their most benign, these grades attempt to define "best practices" and are used as exemplars and resources for improvement in struggling schools. At the opposite end, failing grades bring financial repercussions and school closures that often disproportionately affect already disadvantaged families.

2.2. Accountability System Design and Validity

This trend of increased emphasis placed on "accountability systems" poses a significant measurement challenge: What is school quality? This question is especially thorny due to the myriad interests represented by the wide variety of stakeholders. Such systems are expected to be accurate enough to reliably identify quality or low-performing schools, yet transparent and understandable enough to engage non-experts; precise enough to allow improvements to be tracked over time, yet not too onerous to implement for districts, educators, and students; and in all of this to fully encapsulate the elusive meaning of "quality education" in a way that is

comparable and fair to both an urban school in a low socioeconomic environment and a suburban school with mostly professional parents.

In a review of state accountability systems, Murray and Howe (2017) find common themes for the sixteen states that have implemented systems that assign schools letter grades A-F to denote quality. The authors classify the mechanisms by which these systems are assumed to work into two categories: "bureaucratic accountability" and "market accountability." Most often, the justifications given for these systems are reliant on arguments based in market accountability: letter grades are clear and easy-to-understand, which empowers parents and students to make better decisions. The calculations of these grades are entirely or almost entirely determined by some weighted average of standardized test current-year performance and standardized test score growth. The authors note, however, that they could find no peer-reviewed studies of the internal (construct) validity of such systems.

Figlio and Rouse (2006) find two mechanisms, which would fall into the "market accountability" category, by which receiving a failing grade within such systems can improve the educational output of schools. One is that increased competition—the particular system they describe gave students vouchers to transfer out of low-rated schools to higher-rated schools—forced schools to improve or go out of business. They also suggested that the social stigma attached to a failing grade motivated schools to improve. Based on an examination of performance gains in Florida in response to accountability pressures, they conclude that the stigmatizing effect is the primary motivator for increased performance in failing schools.

This project focuses specifically on one such accountability system for illustration: the New York City School Progress Report (NYCPR). In many ways this particular system pre-emptively implemented the changes now mandated by the ESSA. The NYCPR assigns every

school an A-F grade, based on a panel of measurements. This system was first piloted in 2006 under Chancellor Joel Klein, and implemented city-wide in 2007. The Progress Report has evolved from its initial conception in 2003, and in many ways presaged the changes included in the ESSA. Included are not only measures based on standardized test scores, but also survey results from parents, students and teachers about the subjective aspects of the learning environment. There are also adjustments to account for differences in school contexts—the idea being, in a district as diverse as New York City, the fairest comparison of school quality is among demographically-similar schools. The system is also in a state of constant change, as each year modifications are piloted and tested.

However, a common criticism of the Klein administration was that there was little space for public input during the development process of policy. Gyurko and Henig (2010) describe how "working groups made decisions behind closed doors in a manner reminiscent of nineteenth century progressive reformers designing a system 'for the people but not by the people'" (p. 95). In fact, Hill (2011) notes that this may have been an intentional strategy to avoid the "politics of paralysis." Peck (2014) compares the efforts of the Klein administration with an earlier, failed implementation of an accountability system in New York in the 1970s, and points to this emphasis on speed and reliance on a small group of experts as one of the key reasons for the success of this current iteration.

In the current political environment, validity of an accountability system often seems synonymous with palatability. As the development of the Progress Report occurred "behind closed doors," there has been little public examination on its educational successfulness; that is, the correctly identification of failing and exemplary schools. This may be an example of a problem of circular definitions—there was no standard way of measuring failing or exemplary

schools before this system existed. This does not mean, however, that there is no alternative but to accept the system in its entirety. In fact, there has been much debate over the merits of the Progress Reports, but mostly in non-research settings such as the pages of *The New York Times* (e.g. Gootman & Medina, 2007). It is into the midst of these tensions that this project aims to propose an additional tool in the arsenal to be used in the development in such systems. By providing a statistical framework within which a school quality measurement system can be developed, it allows for the assessment of these varied demands. This work, then, proposes to inject a statistical perspective into the conversation on school accountability system design, by creating a statistical parallel to the Progress Report and using this as a basis for evaluation.

Chapter 3. Statistical Background

The methods in this proposal are at the intersection of several strands within the statistical literature common to social science research. This chapter provides background for three methodologies: propensity score matching, second-order latent factor growth modeling, and Bayesian estimation. But before discussing the statistical methodologies, an important distinction in modeling approaches is presented.

3.1. Emergent Factors and Latent Factors

This project proposes a methodology that is appropriate for examining the effect of an intervention on a construct that is not directly observable but changing over time. Because such a construct is not directly observable, its value is inferred through the observation of other observable or manifest variables. However, it is critical to define the nature of this construct. There are two different types of constructs that cannot be directly observed, and the distinction between the two has important technical and theoretical implications.

The first type is an "emergent factor," a simple example of which would be "net wealth." The net wealth of an individual cannot be directly observed. Instead, it is determined by the sum of a set of observed assets—e.g. bank account balances, home equity, investment holdings—and less a set of observed deficits—e.g. credit card debt, mortgage balances, auto loans. A change in any of these observed variables corresponds to an exact change in net wealth. Net wealth as an "emergent factor" then can be viewed as an effect completely determined by these observed variables.

The second type is a "latent factor," the classic example of which is general intelligence or IQ. Intelligence cannot be directly observed; instead it is often measured through a series

items on a test. These items attempt to give insight into the intelligence of the subject through assessing a variety of domains, such as problem solving, synthesis, visualization, etc. However, the intelligence of the subject is not determined by the number of items correctly answered on the test through any means. That is, if the test-taker were given all the correct answers, a perfect score might indicate the ability to copy accurately, but it would not make the test-taker a genius. Thus, intelligence is viewed as a cause of the number of items answered correctly on the test—potentially one among many—and not an effect.

This distinction between emergent and latent factors is subtle but important. The direction of causality has implications on statistical modeling and its misspecification can lead to errors in estimates and inference (e.g. Bollen, 2002; Bollen & Lennox, 1991; Cohen, Cohen, Teresi, Marchi, & Velez, 1990). Undergirding this project is the argument that School Quality is more appropriately viewed as a latent factor, rather than an emergent factor. This can best be understood by a thought experiment: if a school artificially inflated its students' test scores by cheating, would it be of higher "quality?" If School Quality is similar to net wealth, it is determined by the value of the outcome, not the method by which those outcomes are achieved. Currently this is the perspective taken by almost every school accountability system. However, if it is more similar to intelligence—that is, a latent factor—then similar methodologies used to measure such unobservable constructs can be leveraged here to design better School Quality measures. Some applicable statistical methodologies are described in the following sections.

3.2. Propensity Scores

At their core, school accountability systems are interventions aimed at accomplishing one thing: improving school quality. In many social science fields, estimating the effectiveness of an intervention is difficult because one of the most common methods—using randomized controlled

trials to adjust for pre-existing differences between groups—is infeasible. Propensity score methods, a set of statistical techniques that approximate equivalent groups across all measured covariates, have grown in popularity as a way to account for such differences in non-experimental, observational settings. Propensity scores were initially proposed in a seminal paper by Rosenbaum and Rubin (1983b) as a way to approximate a randomized control trial using observational data. Since then, their use has grown increasingly popular in social science research (Thoemmes & Kim, 2011).

Originally, Rosenbaum and Rubin proposed propensity scores for binary treatments in a regression context. Many researchers have built upon and extended this framework into other contexts. Propensity-based methods have been generalized to multivalued, ordinal, and continuous-type treatments (Hirano, Imbens, & Ridder, 2003; Imai & van Dyk, 2004; Robins, Hernán, & Brumback, 2000); or to frameworks such as structural equation modeling (Hoshino, Kurata, & Shigemasu, 2006). This allows for more sophisticated models that can better reflect complex theories.

The propensity score is a function of the observed covariates—a "balancing score"—such that, conditional on this score, treatment assignment is independent of these covariates. The propensity score is defined within the context of the Neyman-Rubin model for potential outcomes (Rubin, 1974; Splawa-Neyman, Dabrowska, & Speed, 1990). This states that for any individual i, there exists a value of the observation $Y^T(i)$ of the outcome in the treatment condition and $Y^C(i)$ of the outcome in the control condition. Only one of these two outcomes is actually observable; the other is a theoretical counterfactual—what the observed outcome would have been under the opposite assignment. Thus, given treatment $T(i) = 1$ if assigned to the

treatment condition and $T(i) = 0$ if assigned to the control condition, the realized outcome can be given by the equation:

$$Y(i) = T(i)Y^T(i) + \big(1 - T(i)\big)Y^C(i).$$

Using this definition, the average effect of the treatment is given by:

$$ATE = E(Y^T - Y^C)$$

which measures the expected impact of the treatment in the population.

However, changing the location of the estimate may be appropriate depending on the research question. For example, instead of the average effect in the entire population, an investigator might be interested in the impact of treatment on units similar to those which received treatment. This effect is termed the average effect on the treated (ATT), given by

$$ATT = E(Y^T|T = 1) - E(Y^C|T = 1).$$

Alternately, the average effect on the untreated (ATU) measures the impact of treatment if it had been given to those units which did not receive treatment, given by

$$ATU = E(Y^T|T = 0) - E(Y^C|T = 0).$$

In a randomized control trial, these expectations could be directly calculated because there is no expected difference in confounding characteristics, by nature of the randomization. However, in an observational setting, the two groups could differ on a variety of covariates that may cause confounding effects. The challenge is that, in a typical observational setting, there are myriad covariates that could potentially affect treatment assignment. The difficulty of creating similar comparison groups increases exponentially with the number of dimensions. Propensity scores offer a way to reduce the dimensionality of the problem, by collapsing the covariates into a single dimension.

The propensity score models the probability of each individual having been assigned to the treatment or control condition, based on these covariates. That is,

$$e(z_i) = P(T_i = 1|z_i)\,,$$

where z_i is the set of covariates that influence group assignment for individual *i*. Rosenbaum and Rubin show that, conditional on estimates of this score, the assignment to treatment is theoretically independent of the covariates. That is,

$$T \perp z|\hat{e}(z).$$

Rosenbaum and Rubin also show that, conditioned on such a balancing score, the outcomes are conditionally independent of covariates, thus addressing the concern for confounding, at least due to measured covariates. The treatment effects can then be calculated according to the following equations:

$$ATE = E(Y^T - Y^C) = E_{e(z)}[E(Y^T|e(z)) - E(Y^C|e(z))],$$

where the outer expectation is over the distribution of $e(z)$. Similarly,

$$ATT = E_{e(z)|T=1}[E(Y^T|e(z)) - E(Y^C|e(z))], \text{ and}$$

$$ATU = E_{e(z)|T=0}[E(Y^T|e(z)) - E(Y^C|e(z))]$$

Once the propensity score is estimated, usually in a logit-type regression model, there are three common ways in which the conditioning is accomplished: matching, stratification and weighting. Matching draws samples from each treatment condition by selecting based on similarity in propensity score. There are a variety of methods by which this selection can be performed, but this project focuses on *optimal matching* (Rosenbaum, 1989). The *optimal matching* algorithm creates the two samples by minimizing the total difference in propensity

score between matched pairs. That is, the two subsamples will have the property that each treatment unit will have a corresponding control unit, and the sum of absolute differences between the propensity scores of each pair will be the minimum possible while using the maximum number of pairs possible. Within this algorithm there are additional nuances that can be tweaked. A "caliper" can be defined, which is an upper limit on the acceptable distance between propensity scores of matches. This disallows the matching of a treatment unit with a control unit that is too dissimilar. Additionally, each treatment unit can be matched to multiple control units; or multiple treatment units can be matched to the same control unit, in order to improve utilization of the available data.

While matching is the focus of this study, a short description of the other two methods is given here for completeness. Weighting—often referred to as inverse propensity score weighting or covariance adjustment—also creates two balanced groups. However, instead of selecting only a subset from each condition, the full set of available units is included. During estimation, units are weighted to create comparable groups by accounting for each unit's relative contribution. The third method, stratification, separates the propensity score into ranges, that then divide the entire sample into segments, or strata. These strata are assessed for covariate balance. The model is then estimated within each strata, and the results combined as a weighted average to produce the final effect estimates.

There are several assumptions made in using propensity scores. The first is given by:

$$\boldsymbol{T} \perp (\boldsymbol{Y^C}, \boldsymbol{Y^T}) | \boldsymbol{Z}$$

This assumption, commonly known as the Stable Unit Treatment Value Assumption or SUTVA (Cox, 1958; Rubin, 1978), requires that the treatment assignment and the potential outcomes are

conditionally independent on the measured covariates. That is, the response of a unit is conditionally independent on the treatment of other units. The second assumption is given by:

$$0 < \Pr(\boldsymbol{T} = t|\boldsymbol{Z}) < 1$$

This assumes that there is a non-zero probability of either treatment condition for any given set of values of covariates. These two assumptions together are known as strongly ignorable treatment assignment. Given these, Rosenbaum and Rubin show that the propensity score can act as a balancing score, providing unbiased estimates of average treatment effects.

Before any treatment estimates are done, however, checks should be done for assumption violations. The balancing property of the propensity score can be checked by examining differences in distributions of covariates between the resultant treatment and control groups (e.g. Hansen, 2004; Hansen & Bowers, 2008; Rosenbaum & Rubin, 1983a). The requirement of non-zero probability for either treatment or control can be checked by examining the distributions of propensity scores in the resultant treatment and control groups, to ensure that they share regions of common support; that is, to check that there are no selected units whose propensity score is far outside the distribution of the opposite group (Imai, King, & Stuart, 2008; King & Zeng, 2006).

3.3. Second-order Growth Modeling

Structural equation modeling (SEM) provides a framework to model substantive theories that involve variables that cannot be directly measured, such as intelligence or curiosity. It has become increasingly popular for applications in fields as wide-ranging as education and economics to medicine and psychology, as it allows the researcher to explicitly model and test causal theories. The framework itself is also flexible enough to emcompass common statistical techniques such as ANOVA or ANCOVA and also more complex techniques such as item response theory modeling or mediation analysis.

This project focuses on one specific application: second-order growth models (SGM), also known as "curve-of-factors" models (McArdle, 1988). These are an extension of first-order latent growth models (FGMs), more commonly referred to as latent growth models or latent growth curve models (McArdle & Epstein, 1987; Meredith & Tisak, 1990), widely used within longitudinal research. This makes it a natural framework to investigate questions in a wide variety of fields (see Duncan & Duncan, 2009 for an overview).

SGMs and FGMs both represent and measure change over time; but as opposed to the more common FGMs, which examine change in a single manifest variable or a single composite based on multiple manifest variables, SGMs examine change in a theoretical latent factor that is not directly observable. This latent factor is measured through multiple repeatedly-measured manifest variables, as in a factor analysis. A SGM then models the change in this latent construct as a second-order factor. Although both models were proposed around the same time, the FGM has become increasingly common whereas the SGM is still relatively obscure. However, there have been a number of authors recently who have advocated for the adoption of SGMs over FGMs (e.g. Chen, Sousa, & West, 2009; Leite, 2007; von Oertzen, Hertzog, Lindenberger, & Ghisletta, 2010).

Geiser, Keller and Lockhart (2013) provide a useful comparison of these two techniques. The authors note that a FGM assumes that the observed score consists of trait and measurement error influences only and that the variance of situation or person-situation interactions is assumed to be zero; the occasion-specific influences are confounded with time-specific random error. In contrast, SGMs explicitly model occasion-specific errors and time-specific errors separately. This disaggregation of error sources should provide more accurate regression coefficients and smaller standard errors (Ferrer, Balluerka, & Widaman, 2008).

The advantages of SGMs over FGMs are not only technical. In treating the construct of interest as a latent factor, Hancock, Kuo and Lawrence (2001) argue that SGMs are more theoretically defensible. FGMs assume the construct to be an emergent factor, which is inappropriate for many applications, including School Quality as argued in Section 3.1. .

The estimation of a SGM can present challenges because of the increasing complexity with the number of freely-estimated parameters. However, several assumptions regarding measurement invariance can drastically reduce the number of free parameters, as well as improve the interpretability of the results. Stoel, van den Wittenboer and Hox (2004) summarize the hierarchy of invariance definitions, with each definition adding an additional set of constraints on the parameters across measurement occasions. Weak factorial invariance assumes only that the factor loadings are invariant over time. Strong factorial invariance also assumes that the intercepts for indicators are also invariant over time. Strict factorial invariance assumes that, in addition to factor loadings and intercepts, the residual errors are invariant.

Measurement invariance ensures that the same construct is being measured over time by fixing the definition of the construct. There is some debate whether all these invariance assumptions are necessary for interpretability (Meredith & Horn, 2001; Oort, 2001). In the case where a subset of the constraints is relaxed, this results in "partial measurement invariance." This may be appropriate when there are substantive reasons to suspect that a particular aspect of the construct may not be constant over time. Because the addition of these constraints results in nested models, another advantage to SGMs is that these assumptions are directly testable through χ^2 difference tests (Bishop, Geiser, & Cole, 2015; Chan & Bentler, 1998; Ferrer et al., 2008).

Hancock, Kuo and Lawrence (2001) extend the SGM to a multisample setting, by setting an additional set of constraints on the parameters within each group. The factor loadings and

intercepts for all indicators are set to be equal, maintaining the construct invariance across groups. The invariance assumptions can again be tested by relaxing the constraints and testing the χ^2 difference for significance.

3.4. Propensity Score Matching In Structural Equation Modeling

Pearl (2000) argues that SEM is a natural framework for investigating causal claims and can be viewed as an extension of the Neyman-Rubin causal framework. This makes the application of propensity score methods within a SEM model a theoretically compatible combination. In fact, there has recently been more work investigating this overlap (Hoshino et al., 2006; e.g. Kaplan, 1999; Leite, Sandbach, Jin, MacInnes, & Jackman, 2012; Saarela, Stephens, Moodie, & Klein, 2015). However, considering the wide range of applications for SEM, very little of this potential area has been explored.

The intersection of these two methodologies presents a unique difficultly in properly accounting for uncertainty. In the traditional frequentist regime, uncertainty due to matching techniques is often misstated because the propensity score is treated as a fixed quantity in the outcome stage (Gelman & Hill, 2007). Adjusted standard errors are most commonly estimated through bootstrap methods, but even this is rare. Thoemmes and Kim (2011), in a systematic review of techniques used in propensity score literature, found "no explicit mentioning of standard errors that were adjusted for the matched nature of the data… However, we found several studies (12; 14.0%) that reported using bootstrap standard error" (pg. 108), despite few recommendations in the literature (e.g. Tu & Zhou, 2002). However, Abadie and Imbens (2008) found that bootstrap standard errors may not be valid for many common methods for selecting propensity score matched samples.

3.5. Frequentist versus Bayesian Estimation

Moving from a frequentist perspective to a Bayesian perspective offers a natural way to model and propagate forward uncertainty due to multiple sources—from prior knowledge in the literature or expert input regarding distributions of variables or variable selection for models, and also, as specifically regards this project, from propensity scores.

The estimation of propensity scores in a Bayesian framework is motivated by the need to properly account for uncertainty. This improved accounting of uncertainty can provide a more realistic picture of the reliability of model estimates. Rubin, in a 1984 paper arguing for the adoption of Bayesian frameworks in general, notes that "consumers of statistical answers… almost uniformly interpret them Bayesianly, that is as probability statements about the likely values of parameters" (Rubin, 1984, p. 1156). He argues that modeling in a Bayesian perspective better matches the way in which we talk, think, and apply the results of research, and thus is a more intellectually consistent framework.

Bayesian inference does not just provide philosophical comfort, however; there are also significant substantive benefits in the context of policy decisions. Because in a practical sense, every parameter and variable is treated as an unknown quantity within Bayesian inference, a researcher is forced to explicitly state assumptions about sources of error at every step in model estimation. This provides a natural avenue for prior knowledge or other research to be integrated into the estimation of error itself. Uncertainty from all different sources is carried through the whole process. Thus it provides more realistic quantifications of uncertainty, which are important when providing context and nuance to policymakers and stakeholders.

Moving to a Bayesian perspective gives credibility intervals that are 10% wider and slightly more efficient (McCandless, Gustafson, & Austin, 2009) and standard error estimates

that are more reliable in small samples (An, 2010) compared to traditional frequentist methods. Uncertainty from other aspects of propensity score usage can also be modeled in the Bayesian framework. For example, Kaplan and Chen (2014) show how to use Bayesian model averaging to account for uncertainty in propensity score model selection, improving propensity score prediction and increasing uncertainty estimates.

One criticism of using Bayesian estimation with propensity score methods is that it does not accurately reflect the design of a true experimental study (Rubin, 2008). Because the joint likelihood is estimated simultaneously in the Bayesian framework, "feedback" can pass from the observed outcomes to the treatment assignment through the estimated propensity scores. This violates the intended use of propensity scores as originally postulated, which was to approximate a randomized control trial from observed data. The "feedback" does not only violate the design principles, but it can also introduce bias into estimates (Zigler et al., 2013). As a response to these concerns, two-step "quasi" Bayesian estimation methods have been developed (Alvarez & Levin, 2014; Hoshino, 2008; Kaplan & Chen, 2012; McCandless, Douglas, Evans, & Smeeth, 2010). These remove the potential for problematic feedback by severing the estimation process into two parts.

Providing measures of uncertainty that are realistic and intuitive is of tremendous importance for educational research. Perhaps more than most other fields, stakeholders in education have an extremely diverse range of statistical expertise. As more and more high-stakes decisions are being made on the basis of school quality measures, providing context in the form of credibility intervals allows non-statisticians to have informed opinions as to the wisdom of such decisions.

Although this paradigm is not new—Bayes' Theorem has been known since at least 1812—modern advances in computational methods and power have made Bayesian estimation more accessible. As the computing barrier continues to fall, Bayesian methodologies are becoming more prevalent. However, as they are still less familiar especially within educational policy, there is a trade-off in transparency to the non-expert. This is a non-negligible consideration, recognizing that in order for educational research to be translated into actionable policy, it needs to have both validity and understandability to myriad stakeholders. The improvements in realistic error quantification, then, must be paired with an effort to translate the findings to be accessible to the non-statistician.

Chapter 4. Propensity Score Adjusted Second-Order Latent Growth Models (PS-SGM)

This chapter describes the method proposed for incorporating propensity score matching in second-order latent growth models (PS-SGM).

4.1. Notation

Multi-group Second-order Growth Model

The notation here is an expansion of the LISREL parameterization (Jöreskog & Sörbom, 1993; Song & Lee, 2006; Song, Lee, & Hser, 2008). Consider G independent groups measured at *J* time points. In the following notation, the superscript g indicates group membership, where $g = 1, \dots, G$; and the subscript j indicates the time point, where $j = 0, \dots, J-1$. The observed responses or outcomes are represented by $\boldsymbol{Y}_j^g$, endogenous latent variables by $\boldsymbol{\eta}_\mathbf{j}^{\boldsymbol{g}}$, and the exogenous latent variables by $\boldsymbol{\xi}^{\boldsymbol{g}}$. For clarity, the superscript is assumed, except where explicitly stated. These relationships[1] between these variables are then specified by the following equations:

$$\begin{bmatrix} \boldsymbol{Y_0} \\ \vdots \\ \boldsymbol{Y_j} \end{bmatrix} = \mathbf{diag}\left(\boldsymbol{\Lambda_{y_0}} \cdots \boldsymbol{\Lambda_{y_j}}\right) \begin{bmatrix} \boldsymbol{\eta_0} \\ \vdots \\ \boldsymbol{\eta_j} \end{bmatrix} + \boldsymbol{\epsilon}$$

$$\boldsymbol{\eta_j} = \mathbf{B_j}\boldsymbol{\eta_j} + \boldsymbol{\Gamma_j}\boldsymbol{\xi} + \boldsymbol{\zeta_j}$$

The exogenous latent factors are separated into two parts: one part modeling the latent growth, and the other any other exogenous factors.

---

[1] Latent growth modeling can include both time-varying and time-invariant covariates, as well as non-linear growth. However, these are not within the scope of this project; so to simplify the notation, the model is limited.

$$\boldsymbol{\xi} = \begin{bmatrix} \alpha \\ \beta \\ \boldsymbol{\xi}' \end{bmatrix},$$

where here α represents the initial latent factor and β represents the growth latent factor, and $\boldsymbol{\xi}'$ represents any remaining exogenous factors. This also separates the loading matrix into corresponding parts:

$$\boldsymbol{\Gamma_j} = \begin{bmatrix} \boldsymbol{\Gamma_\alpha} \\ \mathrm{j}\boldsymbol{\Gamma_\beta} \\ \boldsymbol{\Gamma}' \end{bmatrix}$$

Specifications of the error terms are given by:

$$\boldsymbol{\epsilon} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\nu_0} \\ \vdots \\ \boldsymbol{\nu_j} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Theta_{00}} & \cdots & \boldsymbol{\Theta_{0j}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Theta_{0j}} & \cdots & \boldsymbol{\Theta_{jj}} \end{bmatrix} \right),$$

$$\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\kappa}, \boldsymbol{\Phi}), \text{ and}$$

$$\boldsymbol{\zeta_j} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi_j}).$$

Assumptions

The following assumptions of the model are expressed as constraints on the parameters listed above.

(1) *Weak factorial invariance.* Weak factorial invariance states that loadings are static. This is given by the constraints:

$$\mathbf{B_0} = \cdots = \mathbf{B_j} = \mathbf{B}$$

$$\boldsymbol{\Gamma_\alpha} = \boldsymbol{\Gamma_\beta}$$

$$\boldsymbol{\Lambda_{y_0}} = \cdots = \boldsymbol{\Lambda_{y_j}} = \boldsymbol{\Lambda_y}$$

(2) *Strong factorial invariance.* Strong factorial invariance assumes weak factorial invariance, and also states that manifest variable intercepts are static. This is given by the constraints:

$$\boldsymbol{\nu_0 = \cdots = \nu_j = \nu}$$

(3) *Strict factorial invariance.* Strict factorial invariance assumes strong factorial invariance, and also states that unique variances are static. This is given by the constraints:

$$\boldsymbol{\Theta_{00} = \cdots = \Theta_{jj} = \Theta_{\epsilon}}$$

$$\boldsymbol{\Psi_0 = \cdots = \Psi_j = \Psi}$$

(4) *Group invariance.* The construct definitions are consistent between groups. Only unique error variances are allowed to differ.

$$\mathbf{B^g = B}$$

$$\boldsymbol{\Gamma^g = \Gamma}$$

$$\boldsymbol{\nu^g = \nu}$$

$$\boldsymbol{\Lambda_y^g = \Lambda_y}$$

(5) *Uncorrelated residuals.* The unique error variances are all independent; the exception being that errors between the same manifest variable across adjacent time points are allowed to co-vary.

$$\boldsymbol{\theta_{ij}} \begin{cases} \text{is diagonal} & \text{, for all } i = j \\ \text{is diagonal and equal to } \boldsymbol{\theta_{ji}} & \text{, for all } |i - j| = 1 \\ \text{is } \mathbf{0} & \text{, for all } |i - j| > 1 \end{cases}$$

(6) *Uncorrelated factors.* Endogenous and exogenous latent factors are all uncorrelated; the exception being the initial and growth latent factors, which are allowed to co-vary.

$\boldsymbol{\Psi}$ is diagonal.

$\boldsymbol{\Phi}$ is diagonal; except $\phi_{\alpha,\beta}$ is allowed to be non-zero.

Propensity Scores

To use propensity scores to estimate this model, let *T* indicate the treatment (*t = 1* for intervention, or *0* for control) for each unit. This treatment variable is used as the grouping variable above. Finally, $\boldsymbol{Z}$ is the vector of all measured covariates. The propensity score, *e(z),* is defined as,

$$e(\boldsymbol{z}) \equiv \Pr(\boldsymbol{T} = 1 | \boldsymbol{Z} = \boldsymbol{z}).$$

The propensity score is modeled by a logistic regression, given by the equation:

$$logit[e(\boldsymbol{z})] = \boldsymbol{a} + \boldsymbol{bz},$$

where $\boldsymbol{a}$ is the vector of intercepts and $\boldsymbol{b}$ is the vector of slopes for the given covariates. The group assignment then is determined by the treatment assignment: $G_i = T_i$.

Optimal Matching

Propensity scores are used to select two subsamples, one for each treatment condition, which have the theoretical property of approximating a random control trial. These samples are selected by identifying "similar" comparison control units for each treatment unit, based on absolute difference in propensity score. The total absolute difference is minimized across all matched pairs without replacement.

Because the original pool of treatment and control units are usually of different size, which treatment effect is estimated can vary depending on which units are used to select. As often the treatment pool is smaller, by selecting the entire treatment pool and finding nearest matches means that the estimated treatment effect will be an estimate of the average treatment effect on the treated (ATT), rather than an estimate of the average treatment effect (ATE).

4.2. Two-step Estimation Method

The estimation method proposed here is a two-step method that propagates the measurement error from the propensity score into the estimation of the second-order growth model. It is an extension of the methodology proposed by Kaplan and Chen (2012), which was applied in the simple regression context.

Bayesian Method

The first method is similar to what Kaplan and Chen denote *BPSA-2*, which applies a Bayesian approach to both propensity score estimation and measurement model estimation. The logic of this method is as follows.

1) The propensity score model is estimated within the Bayesian framework.
2) From the posterior distribution, $m = 1, \ldots, M$ propensity score estimates are drawn for each of the *N* units.
3) Each of the *M* sets of propensity scores is used to create *M* matched control and treatment sets, using optimal matching without replacement.
4) The *M* matched sets are then used to estimate the multi-group second-order latent growth model within the Bayesian framework.
5) For a given parameter, γ, from set m, $j = 1, \ldots, J$ estimates are drawn from its posterior distribution. Each estimate is denoted $\hat{\gamma}_{m,j}$.
6) The final estimate $\hat{\gamma}$ is given by the following equations:

$$\hat{\gamma} = E(\gamma|T, y, z) = M^{-1}J^{-1}\sum_{m=1}^{M}\sum_{j=1}^{J}\hat{\gamma}_{m,j}$$

$$\hat{\sigma}^2(\gamma) = Var(\gamma|T, y, z)$$

$$= M^{-1} \sum_{m=1}^{M} \sigma^2(\gamma_m) + (M-1)^{-1} \sum_{m=1}^{M} \left(\mu(\gamma_m) - M^{-1} \sum_{m=1}^{M} \mu(\gamma_m) \right)^2$$

where $\mu(\gamma_m) = J^{-1} \sum_{j=1}^{J} \gamma_{m,j}$

Maximum Likelihood Method

The second method uses fully ML estimation methods. Although Kaplan and Chen suggest an intermediate Bayesian approach, where the propensity score is estimated in a Bayesian framework, and then the measurement model is estimated in a ML framework, this mixture of perspectives may be too complicated for the average practitioner. Thus, in order to maximize accessibility, a fully ML version is proposed here, with the logic as follows.

1) From the full dataset, $m = 1, \ldots, M$ bootstrap samples are drawn, maintaining the same proportion of treatment and control units.
2) On each of the *M* bootstrap samples, the propensity score model is estimated in the ML framework.
3) Using each of the *M* sets of propensity score model parameters, *M* propensity scores are estimated for each unit in the original full dataset.
4) Each of the *M* sets of propensity scores is used to create *M* matched control and treatment sets from the original dataset, using optimal matching without replacement.
5) The *M* matched sets are then used to estimate the multi-group second-order latent growth model within the ML framework.
6) For a given parameter, γ, the final estimate $\hat{\gamma}$ is given by the following equations:

$$\hat{\gamma} = M^{-1} \sum_{m=1}^{M} \hat{\gamma}_m$$

$$\hat{\sigma}^2(\gamma) = M^{-1}\left(M^{-1}\sum_{m=1}^{M}\hat{\sigma}_m^2(\gamma) + (M-1)^{-1}\sum_{m=1}^{M}\left(\hat{\gamma}_m - M^{-1}\sum_{m=1}^{M}\hat{\gamma}_m\right)^2\right)$$

where $\hat{\gamma}_m$ and $\hat{\sigma}_m^2(\gamma)$ are the traditional ML mean and variance estimates from the m^{th} model estimation.

Chapter 5. Simulation Design

The goal of the simulation study is to explore the efficacy of the proposed model: the propensity-score adjusted second-order growth model (PS-SGM). In particular, the parameter of interest is the treatment effect, here defined as the difference between the latent growth intercept estimates for treatment and control groups.

5.1. Data Generation

We explore the performance of the PS-SGM under several different conditions, following Kaplan and Chen (2012). The simulation parameters are specified to have values similar to those found in the real data application. The data generation is accomplished in four steps.

Step 1) *Outcome generation*

A pool of 10^7 observations is generated as a population using the R packages *lavaan* (Rosseel, 2012) and *simsem* (Jorgensen, Pornprasertmanit, Miller, & Schoemann, 2017), with five indicators across two years of data $y_{0,1}, \ldots, y_{0,5}, y_{1,1}^{g}, \ldots, y_{1,5}^{g}$, where g is a binary indicator of treatment or control. In the second year, both observed and treatment outcomes are generated for all units. The data generation follows the model:

$$\begin{pmatrix} \eta_0 \\ \eta_1^g \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta^g \end{pmatrix} + \Psi$$

$$\begin{pmatrix} \alpha \\ \beta^g \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0.5 + \Delta^g \end{pmatrix}, \begin{pmatrix} 1 & -0.1 \\ -0.1 & 0.5 \end{pmatrix} \right)$$

$$\Psi \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$$

$$\begin{pmatrix} \boldsymbol{y_0} \\ \boldsymbol{y_1^g} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Lambda_y} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Lambda_y} \end{pmatrix} \begin{pmatrix} \eta_0 \\ \eta_1^g \end{pmatrix} + \begin{pmatrix} \boldsymbol{\Lambda_x} \\ \boldsymbol{\Lambda_x} \end{pmatrix} \boldsymbol{X} + \boldsymbol{\Phi}$$

$$\boldsymbol{\Lambda} = \begin{pmatrix} 0.5 \\ 0.7 \\ 1 \\ 1 \\ 1.3 \end{pmatrix}$$

$$\boldsymbol{\Lambda_x} = \begin{pmatrix} -0.4 & -0.3 & -0.2 \\ -0.2 & 0.1 & -0.2 \\ 0 & -0.1 & -0.1 \\ -0.2 & -0.2 & 0.1 \\ -0.2 & -0.3 & -0.3 \end{pmatrix}$$

$$\boldsymbol{\Phi} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{pmatrix} \boldsymbol{\Sigma_{00}} & \boldsymbol{\Sigma_{10}} \\ \boldsymbol{\Sigma_{01}} & \boldsymbol{\Sigma_{11}} \end{pmatrix}\right)$$

$$\boldsymbol{\Sigma_{10}} = \boldsymbol{\Sigma_{01}} = diag(0.6, -0.2, 0.05, 0.1, 0.4)$$

$$\boldsymbol{\Sigma_{00}} = \boldsymbol{\Sigma_{11}} = \begin{pmatrix} 1.0 & 0.1 & 0.2 & 0.05 & 0.3 \\ 0.1 & 0.6 & 0.15 & 0.1 & 0.2 \\ 0.2 & 0.15 & 0.3 & 0.1 & 0.2 \\ 0.05 & 0.1 & 0.1 & 1.1 & 0.05 \\ 0.3 & 0.2 & 0.2 & 0.05 & 0.8 \end{pmatrix}$$

Here, $\Delta^{\mathrm{C}} = 0$ for the control group, and one of the four conditions for the treatment group: $\Delta^{\mathrm{T}} \in \{-1, -0.25, 0.5, 1\}$.

Step 2) *Propensity score generation*

Three covariates z_1, z_2, and z_3 are independently generated with sample size n according to the following distributions.

$$z_1 \sim \mathcal{N}(1,1)$$

$$z_2 \sim \mathrm{Poisson}(2)$$

$$z_3 \sim \mathrm{Bernoulli}(0.5)$$

The true propensity scores are obtained according to

$$e(z) = \frac{1}{1 + \exp\ -(e_0 + 0.2z_1 + 0.3z_2 - 0.2z_3 - 0.4\eta_0)}$$

In this case, an intercept of $e_0 = -2$ was added to the model to shift the expected odds to approximately 1:4, similar to what is observed in school accountability systems. Note that the true propensity score is explicitly dependent on the initial latent factor value, to mimic the process seen in school accountability systems.

Step 3) *Treatment assignment and outcome selection*

The treatment assignment vector *T* is then assigned by generating

$$U_i \sim Unif(0,1)$$

$$T_i\ = \{U_i \leq e_i(z)\}$$

Given the treatment assignment, the final observed values of the corresponding treatment group are retained, according to:

$$(\boldsymbol{y_1})_{\boldsymbol{i}}\ = \left(\boldsymbol{y_1^T}\right)_{\boldsymbol{i}} * (T_i = 1) + \left(\boldsymbol{y_1^C}\right)_{\boldsymbol{i}} * (T_i = 0)$$

Step 4) *Simulation sample generation*

For each of K=1,000 simulation runs, samples are randomly selected from this population pool. Each sample includes 200 treatment units and 600 control units.

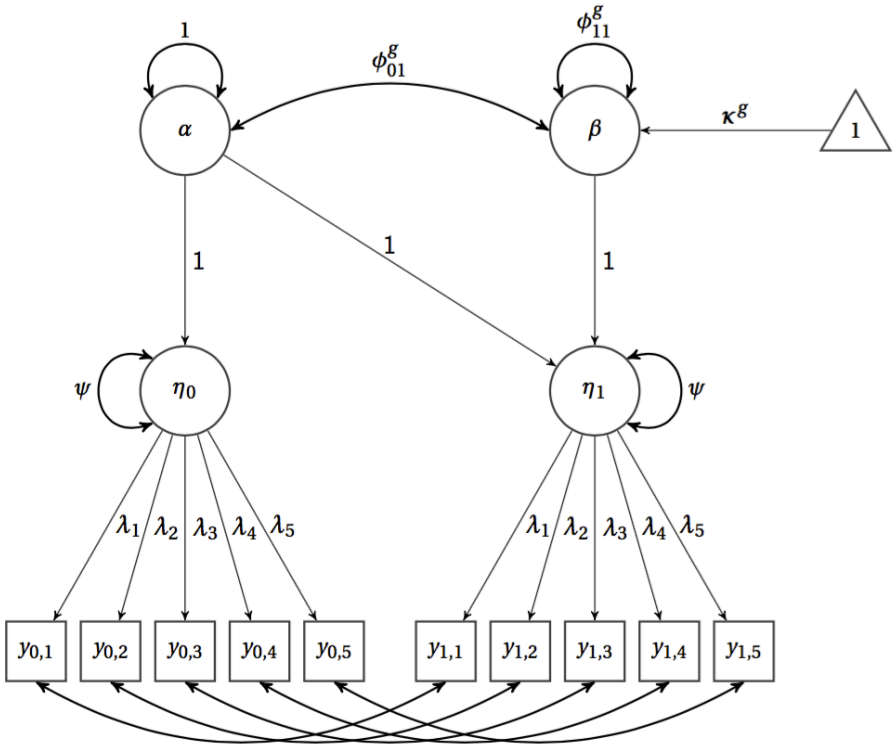

Figure 5.1. Model diagram for estimated PS-SGM.

Note: Omitted from this diagram for clarity are the intercepts and residual variances for each manifest variable. Parameters with the same label are constrained to be equal. The subscript g indicates that the parameter is estimated freely for each group. Fixed parameters are displayed as the fixed value.

5.2. Estimation Models

The four estimation regimes compared here all use the same structural model, shown in Figure 5.1. The first regime uses standard ML estimation as implemented in the *lavaan* R package, from here on referred to as model *ML-full*. The second adds ML-based propensity score matching, using the *optmatch* R package (Hansen & Klopfer, 2006), referred to as model *ML-match*. The estimated propensity scores are calculated using a binomial GLS regression according to the equation:

$$\hat{e}(\boldsymbol{z}) = logit(\widehat{\boldsymbol{b}} \times \boldsymbol{z})$$

Note that this estimated propensity score is mis-specified, as the latent factor score at the initial time point is not available. Matching is then performed on these estimated propensity scores using one-to-one full matching, with no caliper.

The third estimation, model *ML-boot*, maintains the ML estimation regime, but creates a two-step process to account for uncertainty in the propensity score, by drawing j=1,..,J bootstrapped estimates of $\widehat{\boldsymbol{\beta}}$, with J=1,000. The resulting estimated propensity scores, $\hat{e}_{i,j}$, are used in the same one-to-one full matching scheme. These J matched sets are then used to estimate J sets of model parameters, $\hat{\Gamma}_1, \ldots, \hat{\Gamma}_J$. These estimates are combined according to Kaplan and Chen to give the following mean and variance estimates:

$$\hat{\gamma} = \frac{\sum_{j=1}^{J} \hat{\gamma}_j}{J}$$

$$Var(\hat{\gamma}) = \frac{J^{-1} \sum_{j=1}^{J} \hat{\sigma}_j^2 + (J-1)^{-1} \sum_{j=1}^{J} \left(\hat{\gamma}_j - J^{-1} \sum_{j=1}^{J} \hat{\gamma}_j\right)^2}{J}$$

where $\hat{\sigma}_j^2$ is the variance estimate of $\hat{\gamma}$ from the j^{th} bootstrapped propensity score and matched sample.

The last estimation, model *Bayes-match*, is a fully Bayesian regime. The estimation follows a two-step process. First, the $J = 1{,}000$ estimates of the propensity score are achieved through J posterior draws using *MCMCpack* (Martin, Quinn, & Park, 2011). The subsequent

matched sets are then estimated using JAGS using *blavaan* (Merkle & Rosseel, 2015) and *runjags* (Denwood, 2016) in R. Each MCMC run has a burn-in phase of 5,000 iterations and an adaptation phase of 1,000 iterations. Then $m = 1{,}000$ posterior samples are drawn. The final posterior sample mean and variance estimates then are calculated by the following:

$$E(\gamma|T,y,z) = m^{-1}J^{-1}\sum_{i=1}^{m}\sum_{j=1}^{J}\gamma_j(\boldsymbol{b}_i)$$

$$Var(\gamma|T,y,z) = m^{-1}\sum_{i=1}^{m}\sigma^2_{\gamma(\boldsymbol{b}_i)} + (m-1)^{-1}\sum_{i=1}^{m}\left\{\mu_{\gamma(\boldsymbol{b}_i)} - m^{-1}\sum_{i=1}^{m}\mu_{\gamma(\boldsymbol{b}_i)}\right\}^2$$

where $\gamma_j(\boldsymbol{b}_i)$ represents the j^{th} posterior draw of the parameter γ from the model using the matched sample from the i^{th} posterior draw of the propensity score parameters $\boldsymbol{b}$.

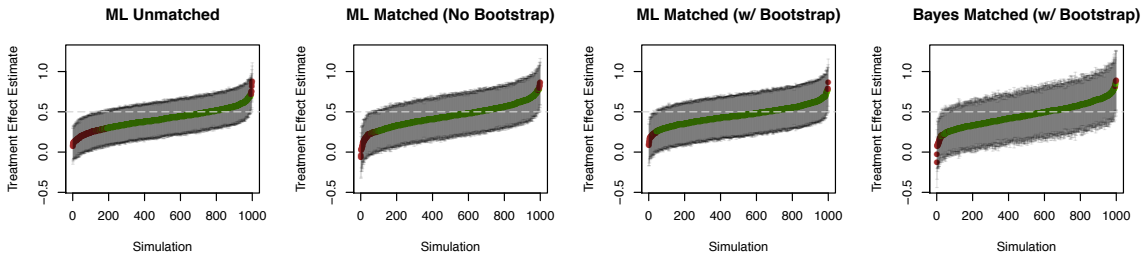

Figure 6.1. Treatment estimates for $\Delta^T = 0.5$ with 95% confidence intervals.

Note: Green points indicate that the true value lies within the confidence interval, whereas red points indicate that the true value is outside the confidence interval.

Chapter 6. Simulation Results

Table 6.1-Table 6.4 show the results of the simulations under the four different treatment effect conditions. The primary constructs of interest here are the Initial and Growth latent factors, α and β, under both control and treatment conditions, as well as the factor loadings. The results show comparisons of biases in estimates and coverage rates under the four estimation methods: *ML-full*, *ML-match*, *ML-boot*, and *Bayes-match*. Note that the variance of the Initial factor in the control group is fixed to unity and the intercept fixed to zero in all methods for the purpose of model identification.

Overall, the estimate biases and coverage rates exhibit a similar pattern across all conditions. Between the *ML-full* and *ML-match* results, the magnitudes in bias either are constant or slightly decrease. At the same time, the coverage rates show marked improvement. Between the *ML-match* and *ML-boot* results, the magnitudes in bias are steady, whereas the coverage rates continue to improve to nearly nominal levels, particularly for the latent factor parameters.

Table 6.1

Simulation mean bias and coverage rates for model parameters with treatment effect $\Delta_T = -1$.

Parameter	ML-naïve	ML-match	ML-boot	Bayes-match
Δ^T	0.019 (90%)	0.077 (87.9%)	0.079 (91.6%)	0.025 (95.5%)
κ^C	-0.037 (88.6%)	0.057 (90.4%)	0.05 (97.8%)	0.09 (96.3%)
ϕ_{00}^T	0.387 (74%)	-0.02 (89.5%)	-0.001 (94.7%)	-0.066 (91.8%)
ϕ_{11}^C	0.218 (77.9%)	0.116 (89.5%)	0.102 (97.6%)	-0.019 (99.9%)
ϕ_{11}^T	0.372 (81.7%)	0.143 (90.8%)	0.113 (95.3%)	-0.025 (100%)
ϕ_{01}^C	-0.106 (65.5%)	-0.093 (91.1%)	-0.079 (93.9%)	-0.01 (99.2%)
ϕ_{11}^T	-0.317 (70.6%)	-0.082 (91.1%)	-0.07 (94.2%)	0.041 (99.1%)
λ_1	0.121 (12.2%)	0.106 (58.8%)	0.111 (52.7%)	0.12 (79.4%)
λ_2	-0.02 (90.2%)	0.025 (94.8%)	0.024 (98.5%)	0.04 (98.7%)
λ_3	0.032 (88.7%)	0.067 (87.9%)	0.064 (94.7%)	0.063 (98.5%)
λ_4	0.096 (51.6%)	0.105 (77.7%)	0.101 (82.8%)	0.03 (99.6%)
λ_5	0.118 (37.7%)	0.133 (71.6%)	0.135 (75.4%)	0.092 (97.8%)

Table 6.2

Simulation mean bias and coverage rates for model parameters with treatment effect $\Delta_T = -0.25$.

Parameter	ML-naïve	ML-match	ML-boot	Bayes-match
Δ^T	-0.034 (90.9%)	0.02 (92.2%)	0.024 (96.9%)	-0.01 (96.1%)
κ^C	-0.042 (87.4%)	0.042 (90.4%)	0.035 (97.8%)	0.091 (96.2%)
ϕ_{00}^T	0.334 (80.7%)	-0.02 (89.3%)	0.001 (94.7%)	-0.06 (92.6%)
ϕ_{11}^C	0.199 (78%)	0.124 (89.2%)	0.111 (97.6%)	-0.022 (100%)
ϕ_{11}^T	0.308 (83.4%)	0.149 (89.9%)	0.12 (95.8%)	-0.029 (99.9%)
ϕ_{01}^C	-0.095 (71.2%)	-0.099 (89.2%)	-0.085 (92.9%)	-0.006 (99.2%)
ϕ_{11}^T	-0.258 (76.2%)	-0.088 (89.2%)	-0.077 (93.4%)	0.044 (98.3%)
λ_1	0.118 (20.5%)	0.115 (55%)	0.113 (54.7%)	0.117 (80.6%)
λ_2	-0.016 (91.4%)	0.012 (94.2%)	0.013 (98.2%)	0.037 (98.9%)
λ_3	0.029 (89.4%)	0.063 (88.6%)	0.06 (95.4%)	0.059 (98.9%)
λ_4	0.093 (56%)	0.114 (74.1%)	0.106 (82.3%)	0.022 (99.7%)
λ_5	0.123 (34.6%)	0.148 (63.9%)	0.153 (64.3%)	0.089 (98.7%)

Table 6.3

Simulation mean bias and coverage rates for model parameters with treatment effect $\Delta_T = 0.5$.

Parameter	ML-naïve	ML-match	ML-boot	Bayes-match
Δ^T	-0.091 (82.3%)	-0.054 (91.1%)	-0.047 (95.4%)	-0.042 (95.9%)
κ^C	-0.031 (90.1%)	0.048 (91.7%)	0.038 (98.2%)	0.088 (96.1%)
ϕ_{00}^T	0.353 (77.1%)	-0.022 (90.6%)	0 (95.1%)	-0.067 (91.6%)
ϕ_{11}^C	0.242 (71.3%)	0.164 (88.6%)	0.153 (96.8%)	-0.017 (100%)
ϕ_{11}^T	0.392 (78.5%)	0.188 (89.2%)	0.159 (95.4%)	-0.026 (99.9%)
ϕ_{01}^C	-0.115 (61.9%)	-0.12 (88.9%)	-0.108 (89.1%)	-0.012 (99.1%)
ϕ_{11}^T	-0.312 (69.3%)	-0.108 (88.9%)	-0.099 (92.9%)	0.041 (99.3%)
λ_1	0.075 (47.1%)	0.079 (65.9%)	0.074 (72.2%)	0.105 (84%)
λ_2	0.01 (91.2%)	0.035 (92.7%)	0.035 (96.3%)	0.041 (98.5%)
λ_3	0.049 (75.1%)	0.088 (79.8%)	0.084 (84.6%)	0.065 (98.2%)
λ_4	0.085 (61%)	0.118 (70%)	0.11 (78.5%)	0.029 (99.4%)
λ_5	0.117 (40.4%)	0.15 (61.2%)	0.154 (61%)	0.095 (97.8%)

Table 6.4

Simulation mean bias and coverage rates for model parameters with treatment effect $\Delta_T = 1$.

Parameter	ML-naïve	ML-match	ML-boot	Bayes-match
Δ^T	-0.115 (79.7%)	-0.092 (87.8%)	-0.097 (91.6%)	-0.067 (94.8%)
κ^C	-0.035 (91.1%)	0.04 (93.3%)	0.037 (98.3%)	0.085 (96.5%)
ϕ_{00}^T	0.287 (84.1%)	-0.002 (91.2%)	-0.002 (95.2%)	-0.072 (92.4%)
ϕ_{11}^C	0.272 (61.8%)	0.191 (86.9%)	0.178 (95.6%)	-0.016 (100%)
ϕ_{11}^T	0.366 (74.6%)	0.189 (88.6%)	0.181 (94.8%)	-0.023 (100%)
ϕ_{01}^C	-0.133 (51%)	-0.134 (88.9%)	-0.123 (86.3%)	-0.016 (98.9%)
ϕ_{11}^T	-0.278 (71.4%)	-0.121 (88.9%)	-0.112 (92.1%)	0.038 (99.1%)
λ_1	0.057 (53.8%)	0.064 (71.3%)	0.064 (74.7%)	0.096 (89%)
λ_2	0.04 (76.7%)	0.049 (89.6%)	0.05 (92.9%)	0.045 (98.2%)
λ_3	0.073 (52.8%)	0.098 (72.6%)	0.098 (76%)	0.07 (97.8%)
λ_4	0.09 (52.3%)	0.112 (71%)	0.113 (74%)	0.036 (99.6%)
λ_5	0.127 (29%)	0.152 (59.3%)	0.154 (58.1%)	0.1 (98%)

Table 6.5
Heywood or ultra-Heywood cases

	Treatment Effect			
Model	-1	-0.25	0.5	1
ML-full	0.0%	0.2%	0.0%	0.1%
ML-match	5.5%	5.0%	3.6%	3.3%
ML-boot	5.6%	5.0%	3.5%	3.0%

Note: ML-boot percentages represent the average across bootstrapped estimates and across simulations.

Comparing *ML-match* and *Bayes-match* results, the magnitudes in bias again tend to decrease, and coverage rates increase. However, with *Bayes-match* the coverage rates are often above nominal levels, suggesting an over-estimation of error variance.

Of particular note is the recovery of the primary parameter of interest, the difference in treatment effects Δ^T. Figure 6.1 shows the treatment estimates with confidence intervals. Both *ML-boot* and *Bayes-match* show better true value recovery rates than the unmatched *ML-full* or single-sample *ML-match* methods. In every case, the coverage rates improve from *ML-full* to the ML-matching methods to *Bayes-match*, which exhibits near-nominal levels of coverage. The magnitude in bias also decreases similarly, except for the case where $\Delta^T = -1$. In this case, it is unclear why *ML-full* outperforms the other methods in bias.

Beyond the benefits in accuracy and coverage, *Bayes-match* offers another advantage over *ML-match* and *ML-boot*. The ML methods occasionally produced Heywood or ultra-Heywood cases—results with zero or negative residual variance estimates. Table 6.5 shows the rates of incidence of such cases. The incidence rates are 3.5% to 5% for both *ML-match* and *ML-boot*, compared to nearly 0% for *ML-full*. The Bayesian estimation eliminates this possibility through the appropriate specification of priors.

Table 6.6
Comparison of average CPU times.

Estimation Method	CPU time
ML-full	2.6s
ML-match	2.7s
ML-boot	1h, 35m, 40s
Bayes-match	5h, 41m, 9s

Bayes-match is much more computationally expensive, however. Table 6.6 shows the average CPU time required for each simulation. Note that the computational cost of *Bayes-match* is nearly four times that of *ML-boot*, even though only $1/100^{th}$ the number of propensity score draws are used. Although the estimation was not optimized—*blavaan* and *JAGS* prioritize flexibility over efficiency—this suggests this difference in computational requirement will be significant.

6.1. Simulation Discussion

The results suggest that propensity score matching methods more accurately recover model parameters for second order latent growth models. The two-step estimation methods, in which the uncertainty due to propensity score estimation is incorporated, achieve near-nominal levels of coverage.

There is not a clear advantage in the choice between Bayesian and ML approaches. The ML approach here offers the benefit of lower computational cost, as well as more established algorithms that provide ease of implementation. The ML-based results, however, exhibit slightly more bias and lower coverage rates than those from the Bayesian approach—potentially related to the Heywood-like cases. The Heywood cases do not appear to be due primarily to model misspecification, as *ML-full* shows almost no such problems. Both *ML-match* and *ML-boot* have

similar rates, suggesting that the cause may be related to the restricted sample due to matching selection.

The Bayesian method is not susceptible to Heywood cases, as they are disallowed by prior specification. However, Bayesian techniques require much more technical expertise, from prior setting to testing for convergence, as well as a much higher computational cost. *Bayes-match* also appears to over-estimate the error variance, leading to higher-than- nominal levels of coverage. Further research is required to investigate whether this can be alleviated by alternate prior specifications. However, this also may be due to the data being generated under a frequentist approach (Kaplan & Chen, 2012; Yuan & MacKinnon, 2009).

Overall, the propensity score methods achieve their purpose to more accurately estimate model parameters, accounting for the impact of exogenous covariates. This is even in the presence of misspecification of the treatment assignment probability. An additional advantage of the propensity score approach is parsimony. In this case, high levels of accuracy and coverage are achieved, despite not explicitly modeling the regressions between the manifest variables and covariates. If these regressions were to be included, the number of additional estimated parameters would increase dramatically as models become more complex. Propensity score matching alleviates the additional associated computational costs and also complexity that might lead to lack of estimation convergence.

Chapter 7. New York City Progress Report Application

The next chapter turns to focus on an application of the proposed framework using propensity score matching within latent factor growth modeling.

7.1. Data

7.1.1. The New York City School Report Card

The New York City Department of Education (NYCDOE) issues an annual Progress Report for every school, as one of three reports used for accountability in New York City schools. These are published by the NYCDOE for each of the over 1,000 schools in New York City serving more than one million students. Here I focus on the 2006-07 and 2007-08 Elementary/Middle School New York City Progress Reports (NYCPR), covering 987 schools. Of these, 893 have complete reported data. Each school is assigned a grade A through F, determined by a weighted average of fifteen measures, intended to capture the complexity of school quality. This project implements the PS-SGM method to provide an alternative to this weighting scheme.

These two years were chosen for several reasons. Firstly, the set of indicators used is consistent across these two years; in subsequent years, the set of indicators is changed. This would violate the assumption of construct invariance, obscuring the interpretation of results. Secondly, the effect of receiving a D or F—the treatment in this model—can be isolated by limiting the analysis to two years. Increasing the number of years analyzed would introduce multiple treatment effects as schools could receive the "treatment" in each year, complicating the isolation of one effect.

7.1.2. School Measures

The fifteen measures used by the NYCPR to calculate the Overall Score are divided into three broad categories: Student Performance, Student Progress, and School Environment.

The Student Performance category comprises four measures. These measures reflect the current-year achievement level of students on the annual New York State standardized exams in Mathematics and English Language Arts (ELA). The first measure is calculated as the percentage of students rated at or above proficient level in ELA. The second is calculated as the median proficiency level of students in ELA. The other two measures are the corresponding calculations for mathematics.

The next category, Student Progress, comprises six measures, based on two-year comparisons of scores from these same tests. The first pair measure the percentage of students making "one year of progress" from the previous year, for mathematics and ELA. The second pair is a similar calculation, limited to students in the lowest-scoring third in each school. The last two measure the average change in proficiency.

The final five measures are within the School Environment category. Four composite scores are based on annual student, parent, and teacher surveys: academic expectations, communication, engagement, and safety and respect. The fifth indicator is the average attendance rate. The aim of these five measures is to capture the essence of school quality not directly associated with standardized test scores.

Note that the details given here are summaries of these measures in order to provide a sense of their substantive meanings. Additional adjustments are performed on each measure to produce a percentile rank city-wide as well as within a demographically similar peer group. The

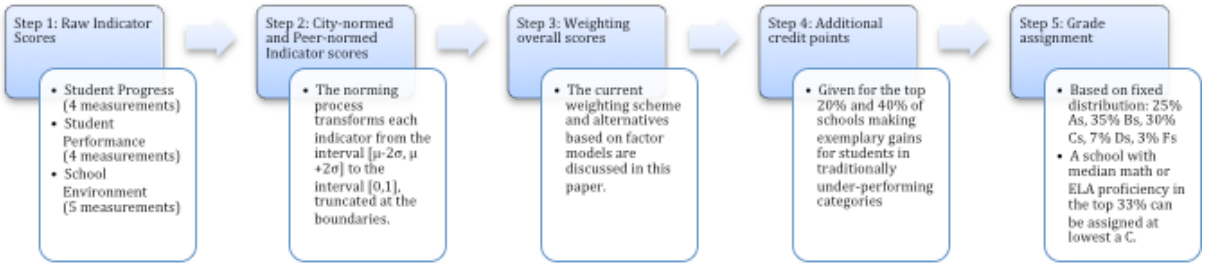

Figure 7.1. The process for calculating the original NYCPR.

entire process is diagrammed in Figure 7.1. For complete details on these calculations and the additional adjustments, see New York City Department of Education (2007).

This preliminary investigation focuses specifically on steps 2 and 3, which calculate a weighted average of the fifteen measures to produce an overall quality score. Adopting an SEM framework derives an alternative weighting scheme for this step of calculating this overall quality score. The tools of SEM can thus investigate the validity of this score. In particular, model fit assessments can give insight into the internal validity of the resultant school quality measures.

7.1.3. Data Transformation

In order to improve algorithmic performance regarding model convergence, the data are normalized according to the following formula, where $i \in \{1, \ldots, 15\}$ indicates which measure and $j \in \{1,2\}$ indicates which year. Here, the $'$ mark indicates the untransformed measure.

$$y_{i,j} = \frac{y'_{i,j} - \bar{y}'_{i,1}}{\sigma(y'_{i,1})}$$

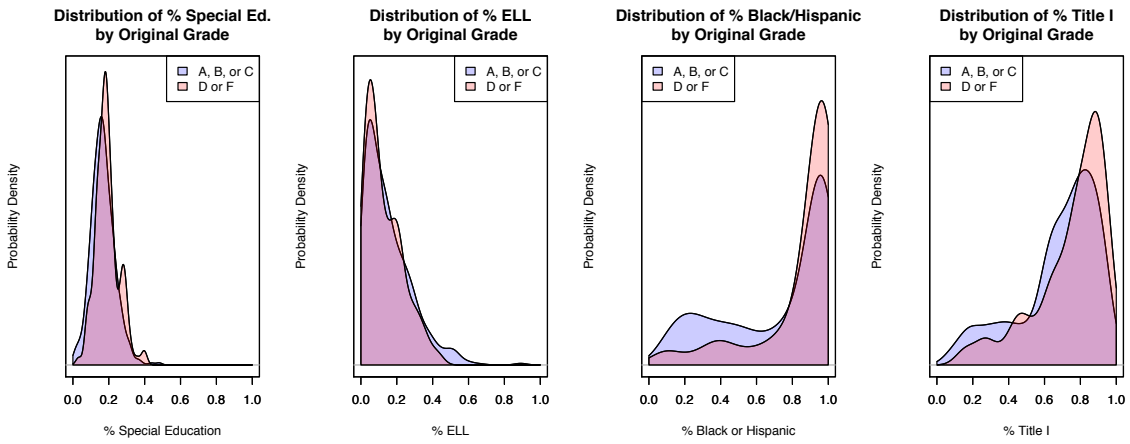

Figure 7.2. Comparisons of demographic distributions between schools assigned passing grades versus failing grades according to the New York City School Progress Report.

7.1.4. School Characteristics

Figure 7.2 shows the density of several demographic characteristics, categorized by the grade received. These four demographics—the percentages of students in each school who are classified as English Language Learners (ELLs), as Black or Hispanic, as qualifying for the Title I Free Lunch program, and as qualifying for special education services—are included here for comparison because they are used in the "peer group index" under the original NYCPR in Step 2 (Figure 7.1). This "peer group" process is intended to facilitate comparisons across similar schools, thus ideally attenuating the influence of demographic differences in measuring school quality.

The distributions of the percentage of ELLs are fairly parallel for schools assigned passing grades versus failing grades. However schools receiving lower grades have noticeably higher concentrations of Black or Hispanic students, of students qualifying for the Title I Free Lunch program, and students receiving special education services. Table 7.1 summarizes the differences between demographics for schools receiving Ds or Fs versus schools receiving As, Bs or Cs in 2006-07.

Table 7.1
Median percentages of demographics

	Citywide	As, Bs or Cs	Ds or Fs	K-S p-value
ELL	11.7%	12.1%	9.5%	0.1135
Black or Hispanic	91.6%	89.7%	96.4%	0.0000
Title I	74.9%	73.7%	80.3%	0.0004
Special Education	16.6%	16.2%	18.5%	0.0002

Note: Results are given citywide (n=987) and subdivided into schools receiving As, Bs or Cs (n=857) and schools receiving Ds or Fs (n=130). Resulting p-values from one-sided Kolmogorov-Smirnov tests between the two subgroups are shown.

The percentages of students who are Black or Hispanic; students qualified for Title I; and students receiving special education services are all higher among schools receiving Ds or Fs, compared with schools receiving As, Bs or Cs. These differences between these two subgroups are statistically significant, according to one-sided Kolmogorov-Smirnov tests. (The difference in percentages of ELLs is the opposite direction, but also not statistically significant.) This suggests that schools that are labeled failing in 2006-07 by the NYCPR are more likely to have higher concentrations of students with these three demographic characteristics.

7.2. Methods

7.2.1. SGM Development

A confirmatory factor analysis was performed using the R package *lavaan* to investigate the model as specified in the original NYCPR. The initial model specifies a pattern of loadings to match the three subdomains in the NYCPR: School Environment, Student Performance, and Student Progress. These three subdomains are then modeled as resulting from a single second-order factor, School Quality. For model identification purposes, loadings were fixed to unity as indicated in Figure 7.3.

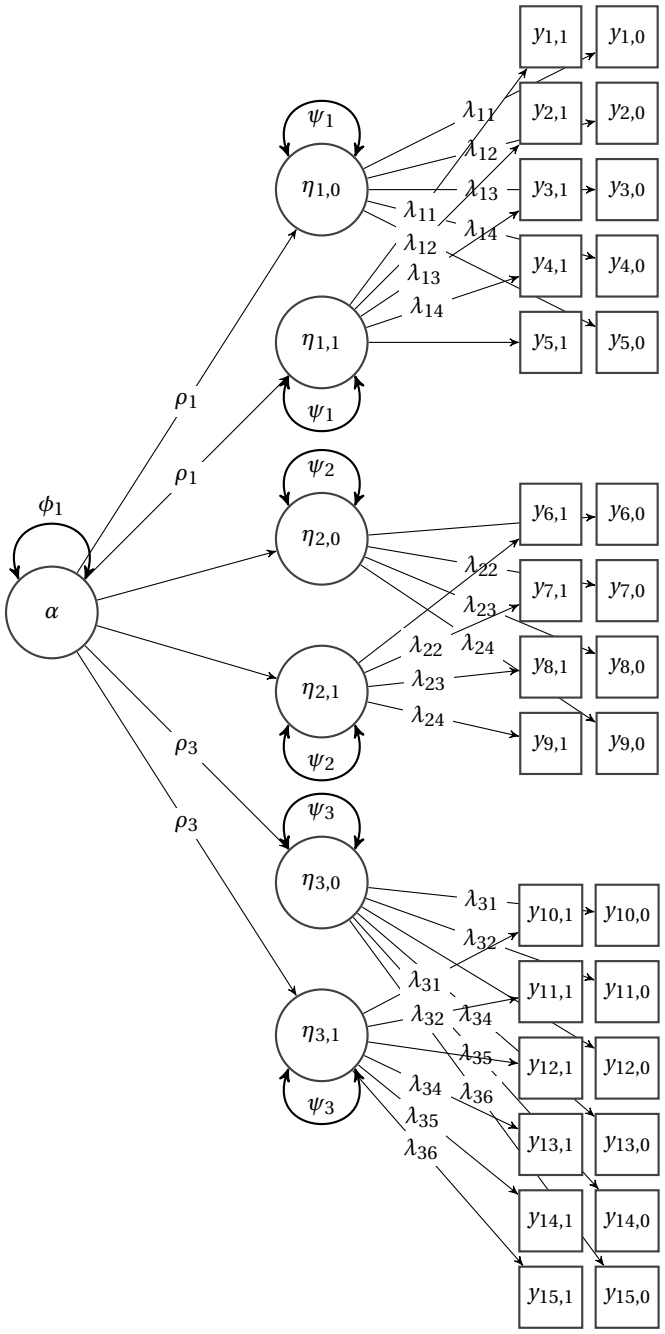

Figure 7.3**.** CFA model representing the original New York City Progress Report structure.

Note: Loadings without labels are set equal to 1 for model identification. Parameters with the same label are constrained to be equal. For clarity, intercepts and unique variances for manifest variables here are omitted.

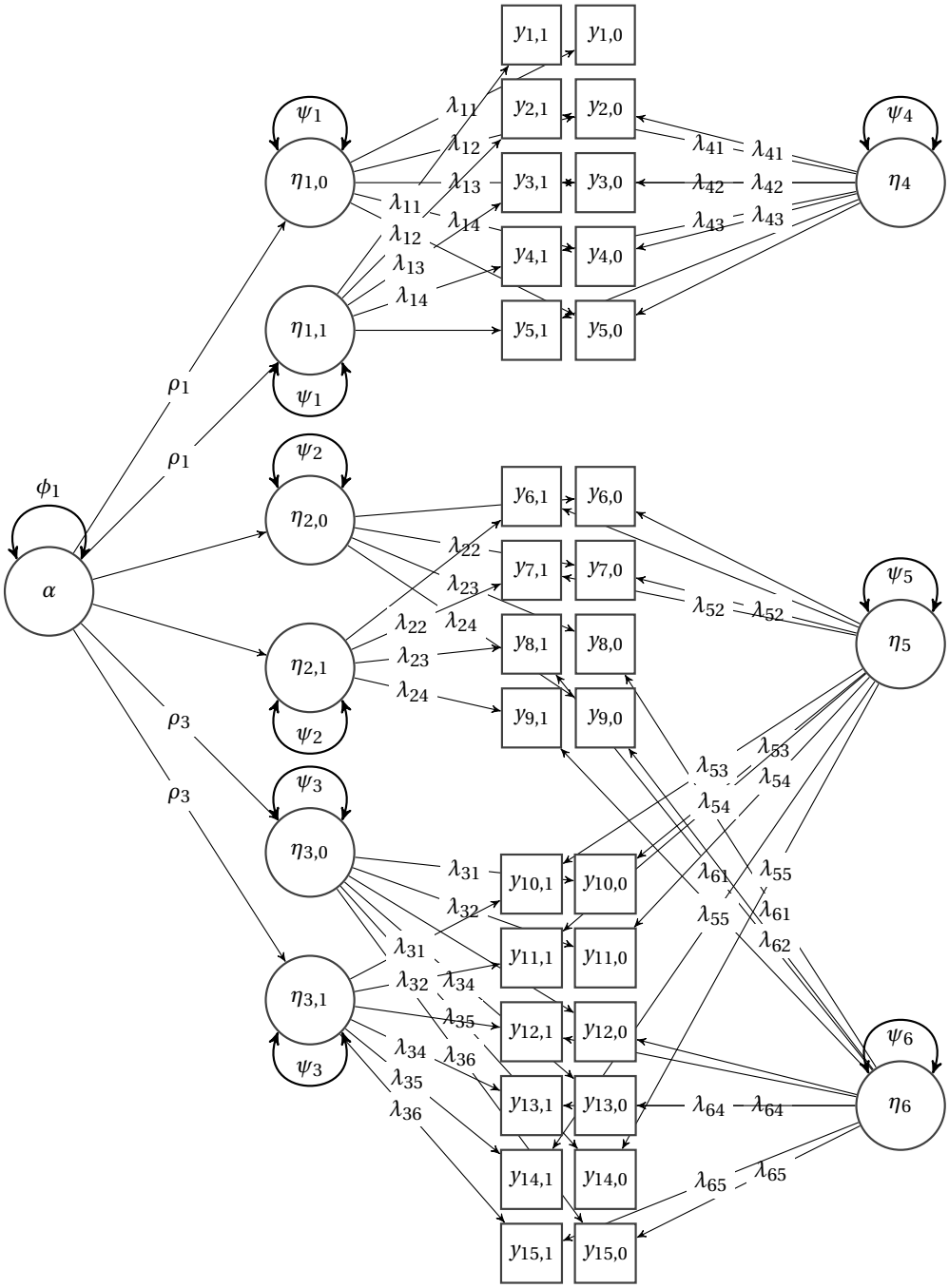

Figure 7.4**.** Final CFA model.

Note: η_4, η_5, and η_6 were successively added. Loadings without labels are set equal to 1 for model identification. Parameters with the same label are constrained to be equal. For clarity, intercepts and unique variances for manifest variables here are omitted.

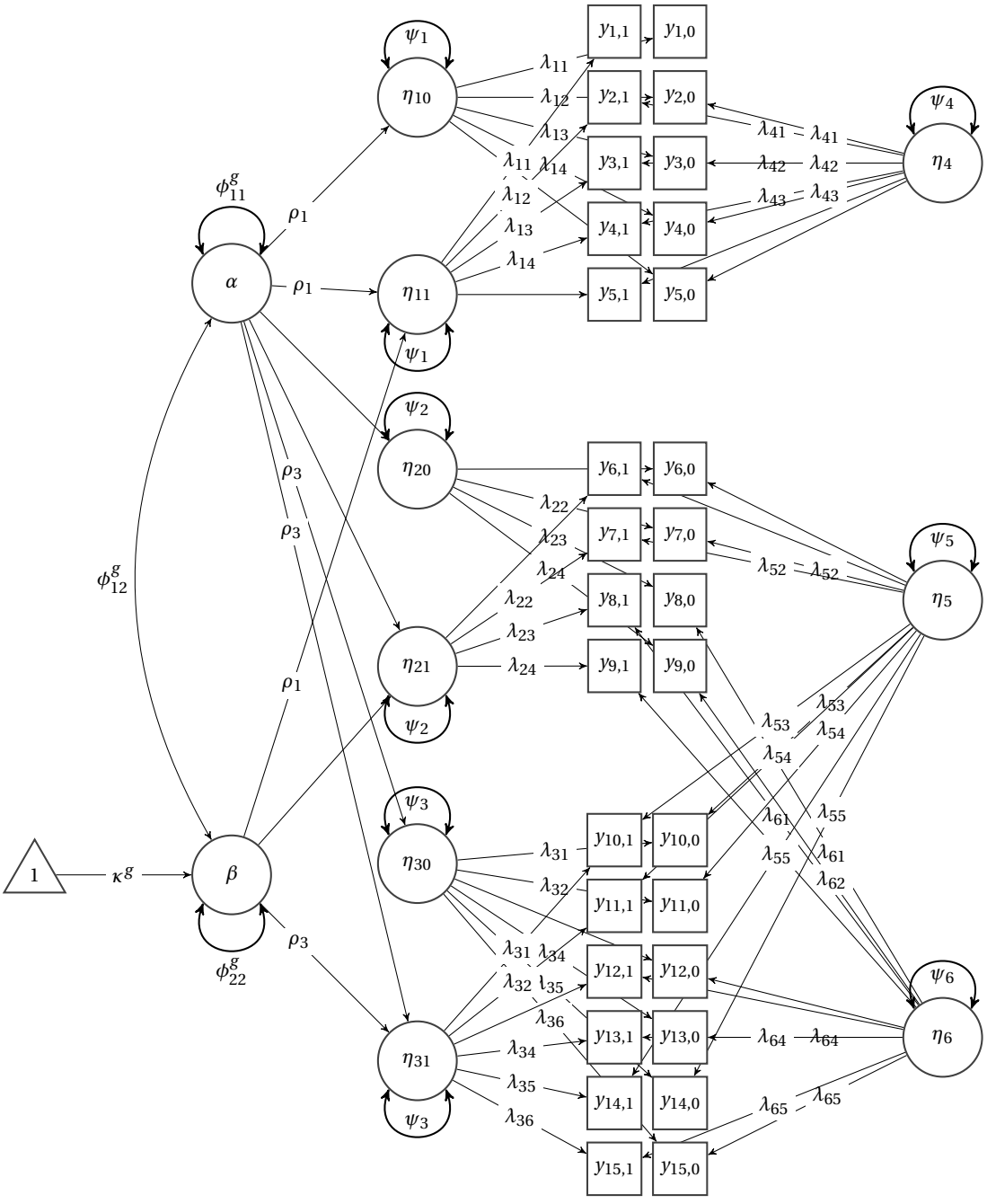

Figure 7.5. Growth model specification.

Note: Loadings without labels are set equal to 1 for model identification. Parameters with the same label are constrained to be equal. For clarity, intercepts and unique variances for manifest variables here are omitted.

Additional method factors are added sequentially to account for the common sources of data collection: the annual student, parent, and teacher surveys; the New York State Mathematics exam; and the New York State English Language Arts exam. The addition of a method factor results in a nested model, which allows the change in model fit to be directly tested. The resulting second-order factor model is shown in Figure 7.4.

The model is then expanded to the SGM shown in Figure 7.5, to incorporate change over time. Strict factorial invariance is assumed, constraining loadings, intercepts, and unique variances for each manifest variable across time, following Ferrer, Balluerka, and Widaman (2008). However, for the two measures measuring the progress of the lowest third of students, there was much criticism during these two years regarding changes to the New York State exams that made it easier for students to score proficient (see, e.g., Kolodner, 2010). The equality constraints between years on the intercepts and unique variances for these two variables are sequentially relaxed, and the change in model fit is assessed.

In order to provide meaningful comparison to the original weighting scheme, correlations between the latent School Quality factor and each measure are reported, re-normalized to sum to unity within each subdomain as well as across the three sub-domains. The re-normalized values indicate the relative importance of each indicator and sub-domain.

7.2.2. Propensity Score Matching Methods

The aim of propensity score matching is the same as that of the peer grouping method used in the NYCPR: to compare schools of similar educational contexts. Because the purpose of this study is to compare the results of the original NYCPR with those from a PS-SGM, the same demographic indices used in the peer index calculation are used here for propensity score estimation. These four indices are described in the Data section.

The propensity score model and matching follow the parameterization and methodology discussed in Chapter 4. The estimation model is given by

$$e(z) = logit(b_{ELL}ELL + b_{TitleI}TitleI + b_{SpecEd}SpecEd + b_{BH}\ Black/Hispanic)$$

7.2.3. Bayesian Estimation

The Bayesian estimation was performed using several R packages: *blavaan* (Merkle & Rosseel, 2018), *runjags* (Denwood, 2016), *rjags* (Plummer, 2016), and *MCMCpack* (Martin, Quinn, & Park, 2011); as well as the Bayesian estimation software JAGS (Plummer, 2015). The estimation was performed using four MCMC chains, with a 25,000 iteration burn-in phase and a 5,000 iteration adaptation phase. Each chain was then sampled 250 times, with a thinning of 400. These values were chosen to minimize autocorrelation and ensure convergence in every sample. Initial values were chosen using the *prior* setting for *blavaan*.

Prior specification can impact the convergence of the MCMC chains. In this case, the priors were used to bound the probability space away from singularities. The priors for all parameters were the default priors set by *blavaan*, with the exception that the priors on Ψ^{-1}, the precisions of the latent variables, were constrained to be greater than 0.05. This prevents the MCMC chains from wandering into flat probability space where the variance estimates are poorly defined. The ranges for each manifest variable are also constrained; the lower and upper bounds are calculated by transforming the maximum and minimum allowable untransformed values for each measure using the same data transformation in Section 7.1.3.

Table 7.2
Distribution of 2007-08 NYCPR Grades

	Elementary	K-8	Middle School
A	10	3	7
B	16	3	6
C	74	23	48
D	220	49	110
F	265	61	147
Total	585	139	318

7.2.4. Grade Calculation

To calculate the model-implied grades, first latent factor scores are estimated. This investigation is limited to the ML bootstrap matching model due to computational constraints. However, since the parameter estimates between the ML bootstrap and Bayesian matching models are consistent, the results are expected to be comparable. The process for this calculation is detailed below.

As the aim of this project was to replace the peer-group and weighting steps in the NYCPR (Steps 2 and 3, Figure 7.1), the remainder of the original process is maintained. The estimated latent factor scores for the final Overall Quality in year 2 are calculated in each ML bootstrapped model. Each model is re-estimated using the full dataset, but constrained to the same model parameters and used to predict latent factor scores. The factor scores are then averaged across the replications. These final scores are re-scaled to the [0,1] interval. Then, additional extra credit score points are added as originally calculated in the NYCPR, and schools ranked according to this resulting score. The grades are assigned according to the same distribution as in the original 2007-08 scores, shown in Table 7.2. Note that the number of schools for which scores are estimated under ML-boot is smaller than the number of 2007-08 scores, as only schools with both years of complete data are included here.

Chapter 8. NYCDOE Results

8.1. SGM Development Results

The resulting changes in model fit from each sequential model change are reported in Table 8.1. A χ^2-difference test directly evaluates the statistical significance in improvement; RMSEA, TLI and CCFI values are also reported. One overall observation is that the model fit in all cases is very poor (RMSEA ranging from 0.27 in the original second order factor model to 0.20 in the final latent growth model). However, each subsequent nested model shows statistically significant improvement in model fit. Because the purpose of this investigation is to provide a statistical model directly comparable to the existing NYCPR system, other possible avenues of model fit improvement are not explored here. Possible alternatives are discussed in the next chapter.

The initial second order factor model, with a loading pattern derived from the NYCPR, exhibits extremely poor model fit (RMSEA = 0.27; TLI = 0.38). This suggests that the factor structure does not explain the data well. By examining the correlations (Table 8.2), it becomes immediately obvious that while the School Environment and Student Performance factors are strongly associated with overall School Quality, Student Progress is almost completely uncorrelated. This suggests that School Quality may not be adequately defined as a single factor, but rather may be multidimensional in nature.

An alternate explanation, however, is that the correlations are masked due to the covariance among indicators due to being derived from the same or similar measurement instruments. To investigate this, I sequentially add additional factors to account for the common methods of data collection discussed in the Data section: the Survey, ELA, and Math factors. With each additional factor, the indices of model fit improve, and the change in χ^2 is statistically

Table 8.1
Fit statistics for each successive nested model

	RMSEA	CFI	TLI	NNFI	DF	$\Delta\chi^2$	
CFA Models							
Second order factor model	0.27	0.36	0.38	0.38	447		
w/ Survey factor	0.27	0.39	0.40	0.40	443	1263.68	*
w/ ELA factor	0.25	0.46	0.46	0.46	438	3156.81	*
w/ Math factor	0.25	0.48	0.47	0.47	433	825.88	*
Latent Growth Models							
Strict factorial invariance	0.24	0.51	0.51	0.51	430	1642.46	*
Relaxed intercepts	0.20	0.66	0.65	0.65	428	6797.34	*
Relaxed errors	0.20	0.66	0.65	0.65	426	87.29	*

Note: The changes in χ^2 are also reported, with * indicating a statistically significant change ($p < 0.001$).

significant, suggesting that the additional factors are significantly improving the explanatory power of the model. With all three method factors, the model fit improves (RMSEA = 0.25; TLI = 0.47). This model is then used as the basis for the second order latent growth model.

Because of the assumption of strict factorial invariance, the CFA model is nested within the SGM. This allows the use of the χ^2 likelihood ratio test to evaluate model improvement. Directly modeling the change over time in a latent growth model leads to an improvement in model fit (RMSEA = 0.24; TLI = 0.51) and a statistically significant change in χ^2.

This assumption of strict factorial invariance may be untenable for two indicators: the percentages of students in the lowest third of each school with one year of growth in ELA and math. To test the validity of these assumptions, the equality constraints on the intercepts and then the unique variances of the two manifest variables are sequentially relaxed. The relaxation of the

Table 8.2
CFA results

		Confirmatory Factor Model			
Indicator	NYCPR	Second Order	w/ Survey	w/ ELA	w/ Math
School Environment					
Attendance	33.3%	8.7%	15.0%	13.7%	15.0%
Academics	16.7%	23.9%	21.6%	21.9%	21.6%
Communication	16.7%	22.5%	20.5%	21.2%	20.4%
Engagement	16.7%	24.8%	22.1%	22.5%	22.0%
Safety and Respect	16.7%	20.0%	20.8%	20.8%	20.8%
Performance					
% Proficient: ELA	25.0%	27.1%	27.1%	24.2%	26.7%
Median: ELA	25.0%	24.7%	24.7%	21.6%	24.4%
% Proficient: Math	25.0%	24.1%	24.1%	27.4%	24.4%
Median: Math	25.0%	24.1%	24.1%	26.8%	24.5%
Progress					
% with 1 Year Growth: ELA	12.5%	8.6%	8.6%	8.6%	6.6%
% in lowest 3rd with 1 yr Growth: ELA	12.5%	39.8%	39.8%	40.1%	39.7%
% with 1 Year Growth: Math	12.5%	10.7%	10.7%	10.6%	11.1%
% in lowest 3rd with 1 yr Growth: Math	12.5%	29.3%	29.3%	29.2%	31.3%
Average Change: ELA	25.0%	7.0%	7.0%	6.9%	6.1%
Average Change: Math	25.0%	4.7%	4.7%	4.6%	5.1%
Overall					
School Environment	15.0%	40.9%	58.0%	57.9%	57.2%
Performance	25.0%	63.3%	45.0%	42.5%	45.5%
Progress	60.0%	-4.2%	-3.0%	-0.3%	-2.6%

intercept constraints leads to the largest χ^2 change per degree of freedom and improvement in model fit (RMSEA = 0.20; TLI = 0.65). The relaxation of the unique variance constraints leads to a much smaller χ^2 change—although still statistically significant—and no change in model fit statistics (RMSEA = 0.20; TLI = 0.65).

The patterns of correlations in the latent growth model results (Table 8.3) offer some insight into what is driving this change in model fit. As previously noted, the CFA model suggested that the Student Progress factor may not fit well with the unidimensional definition of

Table 8.3
Latent Growth Model results

		Latent Growth Model		
Indicator	CFA Model	Strict Invariance	Relaxed Intercepts	Relaxed Errors
School Environment				
Attendance	15.0%	13.5%	13.5%	13.5%
Academics	21.6%	22.2%	22.2%	22.2%
Communication	20.4%	21.1%	21.1%	21.1%
Engagement	22.0%	22.7%	22.7%	22.7%
Safety and Respect	20.8%	20.5%	20.5%	20.5%
Performance				
% Proficient: ELA	26.7%	24.0%	24.0%	24.0%
Median: ELA	24.4%	21.2%	21.1%	21.1%
% Proficient: Math	24.4%	27.9%	27.8%	27.8%
Median: Math	24.5%	26.9%	27.1%	27.1%
Progress				
% with 1 Year Growth: ELA	6.6%	9.3%	27.5%	27.7%
% in lowest 3rd with 1 yr Growth: ELA	39.7%	42.1%	17.3%	16.9%
% with 1 Year Growth: Math	11.1%	9.2%	12.8%	12.9%
% in lowest 3rd with 1 yr Growth: Math	31.3%	28.7%	6.9%	7.1%
Average Change: ELA	6.1%	7.6%	26.5%	26.4%
Average Change: Math	5.1%	3.1%	9.1%	9.0%
Overall				
School Environment	57.2%	52.2%	59.8%	59.7%
Performance	45.5%	33.4%	38.4%	38.3%
Progress	-2.6%	14.4%	1.7%	2.0%

School Quality. Moving to the latent growth framework, however, increases the relative importance of Student Progress to 14.4%. When the intercept constraints are relaxed, however, this importance falls back to 1.7%. This suggests that much of the increase in relative importance is due to the artificial inflation of the two indicators in question in the second year.

This is supported by the large improvement in model fit and the reversion of the relative importance of Student Progress after these two constraints are released. A similar change is not observed with the relaxation of the constraints on the unique variances. The relative importance

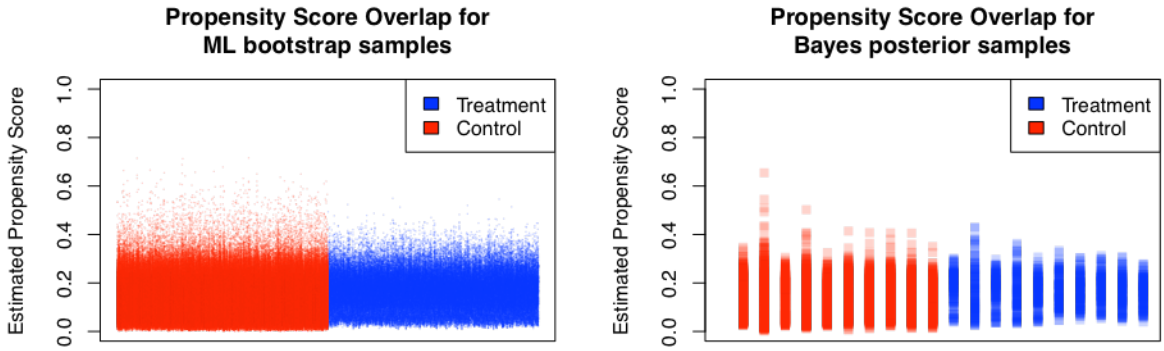

Figure 8.1. Propensity score estimates.

Note: Each set of propensity score estimates, plotted in vertical lines, are based on separate draws of propensity score parameters.

of indicators and factors—as well as model fit indices—barely change. Thus, although the χ^2 test is statistically significant, these constraints are re-imposed to maintain parsimony. Thus the SGM, with partial strict invariance except for the two relaxed intercept constraints, is the basis for estimation in the next step: incorporating propensity score matching.

8.2. Propensity Score Analysis

The results of the assumption checks are shown here. Figure 8.1 shows the propensity score estimates estimated using each bootstrapped sample or posterior draw of the propensity model parameters. The scores show strong overlap between the treatment and control subgroups, suggesting that there is sufficient common support.

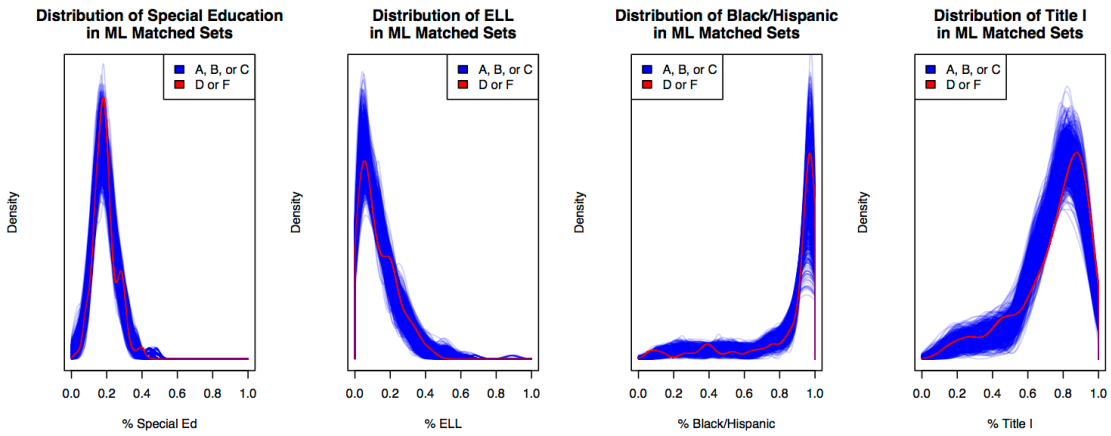

Figure 8.2**.** Comparison of demographics between ML matched sets.

The demographic distributions for the resulting matched samples are shown in Figure 8.2 (for maximum likelihood) and Figure 8.3(for Bayesian estimation). Along all the demographics, the treatment and control groups have similar densities, again suggesting that the propensity score matching has accomplished covariate balance between the two groups.

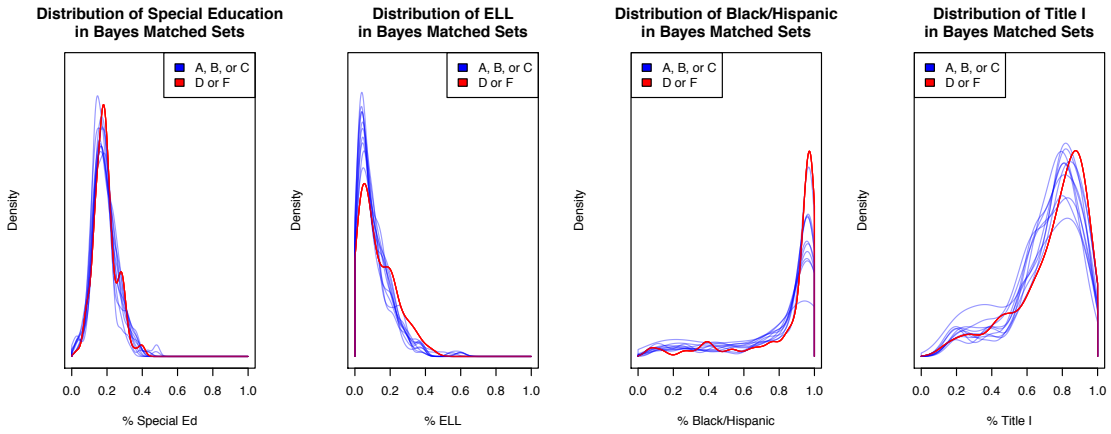

Figure 8.3**.** Comparison of demographics between Bayes matched sets.

Table 8.4
PS-SGM Results

Indicator	NYCPR	ML full	ML match	ML boot	Bayes Match
School Environment					
Attendance	33.3%	13.5% (12.9%, 14.2%)	10.0% (8.6%, 11.3%)	11.8% (11.0%, 13.4%)	11.4% (10.4%, 12.5%)
Academics	16.7%	22.1% (21.7%, 22.6%)	24.1% (23.2%, 25%)	23.2% (22.1%, 24%)	23.3% (21%, 25.8%)
Communication	16.7%	21.1% (20.6%, 21.6%)	22.0% (21.0%, 22.9%)	21.6% (20.5%, 22.5%)	21.8% (19.8%, 23.8%)
Engagement	16.7%	22.7% (22.1%, 23.2%)	25.1% (24.0%, 26.3%)	23.5% (22.1%, 24.4%)	23.5% (20.4%, 26.5%)
Safety and Respect	16.7%	20.6% (20.3%, 20.8%)	18.8% (18.3%, 19.3%)	19.9% (19.6%, 20.5%)	19.9% (18.9%, 21.0%)
Performance					
% Proficient: ELA	25.0%	24.0% (23.7%, 24.2%)	23.1% (22.7%, 23.5%)	23.7% (23.3%, 24.2%)	24.3% (24.1%, 24.6%)
Median: ELA	25.0%	21.1% (20.7%, 21.4%)	20.0% (19.4%, 20.6%)	20.6% (20.0%, 21.3%)	20.9% (20.6%, 21.2%)
% Proficient: Math	25.0%	27.9% (27.4%, 28.4%)	29.9% (28.9%, 30.8%)	29.1% (28.0%, 30.1%)	29.0% (28.4%, 29.5%)
Median: Math	25.0%	27.1% (26.6%, 27.6%)	27.0% (26.1%, 27.9%)	26.6% (25.5%, 27.5%)	25.8% (25.3%, 26.3%)
Progress					
% with 1 Year Growth: ELA	12.5%	28.5% (25.8%, 30.9%)	24.1% (21.6%, 26.5%)	23.7% (21.2%, 26%)	23.8% (22.1%, 25.5%)
% in lowest 3rd with 1 yr Growth: ELA	12.5%	17.7% (15.8%, 19.8%)	16.0% (14%, 18%)	16.0% (13.8%, 18.1%)	16.7% (15.4%, 18.0%)
% with 1 Year Growth: Math	12.5%	12.0% (11.4%, 12.6%)	15.6% (14.8%, 16.4%)	15.7% (14.8%, 16.5%)	15.4% (14.8%, 15.9%)
% in lowest 3rd with 1 yr Growth: Math	12.5%	6.4% (5.2%, 7.6%)	8.3% (6.7%, 9.9%)	9.1% (7.4%, 11.0%)	9.0% (8%, 9.9%)
Average Change: ELA	25.0%	27.3% (24.9%, 29.8%)	23.1% (20.9%, 25.5%)	22.9% (20.4%, 25.3%)	22.9% (21.1%, 24.6%)
Average Change: Math	25.0%	8.1% (6.8%, 9.3%)	12.9% (10.9%, 14.9%)	12.7% (11.0%, 14.6%)	12.2% (11.4%, 13.0%)
Overall					
School Environment	15.0%	59.9% (58.2%, 61.6%)	51.6% (47.7%, 55.6%)	57.0% (53.4%, 61.7%)	55.0% (47.6%, 61.5%)
Performance	25.0%	38.4% (37.1%, 39.7%)	39.5% (36.3%, 42.6%)	34.6% (31.9%, 37.5%)	35.6% (30.4%, 41.3%)
Progress	60.0%	1.8% (-0.4%, 3.8%)	8.9% (3.4%, 14.4%)	8.4% (2.5%, 13.1%)	9.4% (6.2%, 13.1%)

Note: 95% confidence intervals shown in parentheses.

Table 8.5
PS-SGM parameter estimates

Unstandardized Estimates	ML full data	ML matched	ML boot	Bayes match
Initial Variance, Control	0.581 (0.033)	0.486 (0.069)	0.515 (0.074)	0.662 (0.077)
Initial Variance, Treatment	0.756 (0.102)	0.493 (0.072)	0.572 (0.081)	0.701 (0.018)
Initial/Growth Covariance, Control	-0.066 (0.007)	-0.043 (0.017)	-0.042 (0.017)	-0.122 (0.015)
Initial/Growth Covariance, Treatment	-0.093 (0.023)	-0.042 (0.019)	-0.058 (0.021)	-0.111 (0.009)
Growth Variance, Control	0.007 (0.004)	-0.037 (0.012)	-0.015 (0.011)	0.034 (0.001)
Growth Variance, Treatment	0.028 (0.009)	-0.022 (0.013)	0.004 (0.013)	0.049 (0.002)
Growth intercept	0.331 (0.008)	0.379 (0.019)	0.367 (0.019)	0.395 (0.010)
Treatment effect	0.107 (0.021)	0.101 (0.026)	0.120 (0.027)	0.089 (0.010)

8.3. PS-SGM Results

The results of the model estimates based on matched samples are shown in Table 8.4. Five models are reported here for comparison: the original NYCPR; ML estimation on the full dataset (*ML-full*); ML estimation on a single matched dataset (*ML-match*); ML estimation on 1,000 matched samples from bootstrapped propensity score estimates (*ML-boot*); and Bayes estimation on 10 matched samples from posterior propensity score draws (*Bayes-match*). The 95% confidence intervals are also shown, using the adjusted standard errors. Note that the Bayesian intervals are not true credibility intervals, as they were not drawn from the posterior distribution. Instead, they are calculated using normality assumptions, following Kaplan and Chen (2012). I highlight some specific patterns here.

The relative loadings on the three subdomains are markedly different in the PS-SGM models, compared to the NYCPR. As previously noted, the latent growth modeling approach suggests that the original NYCPR vastly overweighted the Progress domain in calculating School Quality. In fact, the loading for the Progress domain is not statistically significant in full dataset. In the matching regimes, however, the loadings are all now significantly non-zero. The relative importance of the Progress domain increases from 1.8% in the full dataset to 8.9% in the matched sample. Both the ML and Bayes PS-SGM estimates give similar results, at 8.4% and

Table 8.6
Model estimation errors

Bootstrapped ML-propensity score estimation errors	
Heywood or Ultra-Heywood cases	93.4%
Non-positive definite model-implied covariance matrix	5.4%

9.4% respectively. This suggests that Progress is still relatively more important to schools at the lowest end of the School Quality distribution. The results also suggest that School Environment is much more strongly associated with School Quality than assumed under the NYCPR. Under all SGMs, the relative importance of School Environment is highest among the three subdomains.

Within each subdomain the patterns of loadings are also informative. Within School Environment, attendance is less important while the measures based on student, teacher and parent feedback are more important. Student Performance remains relatively untouched, although the emphasis on students scoring proficient in math is slightly increased and the median ELA score is slightly de-emphasized. The loadings within Student Progress see the most change. The contribution from the percentage of students showing a year of growth in ELA nearly doubles, whereas the contribution from the average change in math is halved. In general, the estimates between matching methods are consistent, with overlapping confidence intervals.

The estimates of the PS-SGM parameters are shown in Table 8.5. There are several items to note here. The ML estimates on the full dataset exhibit problems characteristic of Heywood cases. The correlation between the initial and growth factors, given by $r_{12} = \frac{\phi_{12}}{\sqrt{\phi_{11}\phi_{22}}}$, is -1.024 for the control group. Similarly for the ML matched samples estimates, the estimated variances of the growth factor are negative for both treatment and control for single-sample matching, and for control for multi-sample matching. Indeed, Heywood-type errors in estimation occur in

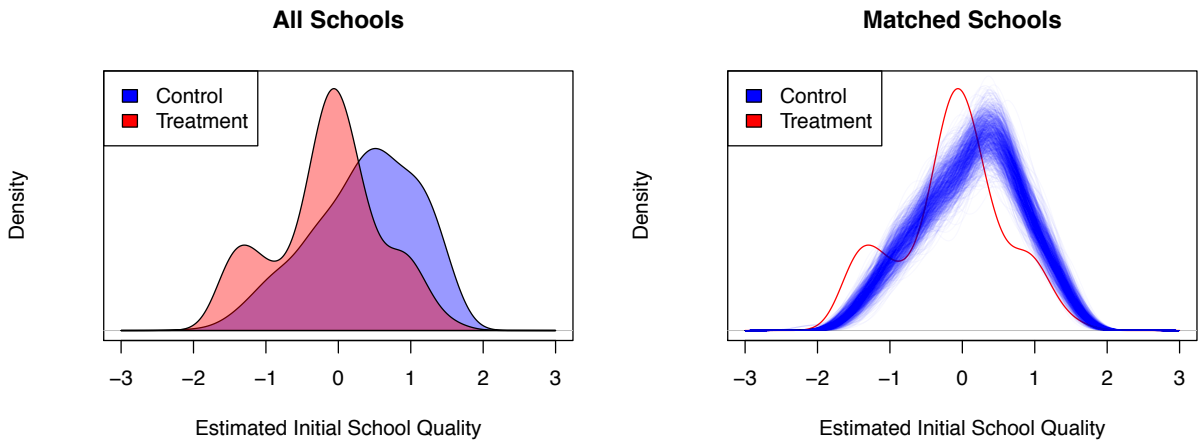

Figure 8.4. Comparisons of estimated densities for imputed initial school quality.

Note: The scores are imputed for treatment and control groups, using the *ML-boot* model. The left panel shows the densities for all schools, whereas the right panel shows trace lines for each matched sample.

almost every bootstrapped sample (Table 8.6). The Bayesian estimates avoid this issue *a priori* by constraints on the prior. However, the presence of these Heywood cases may be another indication of model misidentification—this will be further discussed in the next chapter.

The estimates of the magnitude of average growth for the control group are approximately half of the standard deviation in Initial School Quality for both ML-boot ($\frac{0.367}{\sqrt{0.515}} = 0.511\sigma$) and Bayes ($\frac{0.395}{\sqrt{0.662}} = 0.485\sigma$) matching methods. For the treatment group, the growth in School Quality shows a statistically significant additional increase of $\frac{0.120}{\sqrt{0.515}} = 0.167\sigma$ for ML-boot and $\frac{0.089}{\sqrt{0.662}} = 0.110\sigma$ for Bayes. This suggests that on average, schools improved in School Quality between the 2006-07 and 2007-08 school years; and schools identified as failing in 2006-07 showed larger improvement than those not.

Because the model has the Initial Quality intercept constrained equal to zero in both groups, the estimates of growth are point estimate comparisons for changes in School Quality for

Table 8.7
Comparison of grades from NYCPR and ML-boot.

		ML-boot				
		A	B	C	D	F
NYCPR	A	195	142	24	1	0
	B	140	164	72	22	5
	C	33	64	37	12	3
	D	2	15	12	10	8
	F	2	8	1	1	2

Note: Kendall's rank correlation coefficient $r_\tau = 0.319, p < 0.001$

a school with an estimated Initial School Quality value of zero, analogous to controlling for Initial School Quality in a regression context. It is important, then, to ensure there is sufficient coverage at this point for both groups. Figure 8.4 shows the distributions for the imputed values of the Initial School Quality, based on *ML-boot*. Although the distributions are not identical, there is significant overlap at all points, with the maximum of each distribution near zero for both control and treatment groups. Compared to the estimated density of all control schools, the estimated densities for the matched control schools in each bootstrapped propensity score model appear to overlap more with that of the treatment schools. Considering the simulation results—where mis-specifying the propensity score model by omitting the influence of the initial latent factor still resulted in nominal recovery rates of treatment effect estimates—these provide support for the interpretation of the difference in the Growth intercepts as a causal treatment effect.

8.3.1. Comparison of Resulting School Quality Grades

Table 8.7 shows the distribution of School Quality Grades, compared between those given by the original NYCPR and those derived using *ML-boot*. Fewer than half $\left(\frac{408}{975} = 41.8\%\right)$ of schools are assigned the same grade, and quite a few $\left(\frac{116}{975} = 11.9\%\right)$ are assigned grades two

Table 8.8
Comparison with Federal Accountability ratings

Federal Accountability	NYCPR A	B	C	D	F
Good Standing	263	265	100	21	12
Needs Improvement	50	47	16	6	0
Corrective Action	11	25	14	6	0
Restructuring	38	66	19	14	2

Federal Accountability	*ML-boot* A	B	C	D	F
Good Standing	327	240	70	16	8
Needs Improvement	20	64	24	7	4
Corrective Action	18	26	6	5	1
Restructuring	7	63	46	18	5

Note: Kendall's rank correlation coefficients for NYCPR ($r_\tau = 0.087, p = 0.002$) and *ML-boot* ($r_\tau = 0.340, p < 0.001$).

or more letter grades apart. The Kendall's rank coefficient ($r_\tau = 0.319, p < 0.001$) suggests that, while related, the two grading systems are not interchangeable.

One of the fundamental issues at hand is that there does not exist a pre-defined theory of what constitutes School Quality. To examine external validity, therefore, I compare these grades with two other scales designed to accomplish a similar purpose in identifying schools in need of improvement: the Federal Accountability Status and the New York City Quality Review Score. The Federal Accountability Status of a school is determined by its achievement of Annual Yearly Progress (AYP), the benchmark as defined under the No Child Left Behind Act. A school that falls short of its designated target in successive years moves down through the levels, from being identified as Needs Improvement, to mandated Corrective Action, until finally it is put in the process of Restructuring. Thus, those schools that continually do not achieve their targets progressively face harsher and harsher sanctions.

Table 8.9
Comparison with School Quality Review scores

	NYCPR				
School Quality score	A	B	C	D	F
Outstanding	13	2	0	0	0
Well Developed	268	249	83	15	7
Proficient	80	146	58	32	6
Underdeveloped with Proficient Features	1	6	6	0	1
Underdeveloped	0	0	1	0	0

	ML-boot				
School Quality score	A	B	C	D	F
Outstanding	11	3	1	0	0
Well Developed	291	250	63	14	4
Proficient	69	131	80	28	14
Underdeveloped with Proficient Features	1	7	2	4	0
Underdeveloped	0	1	0	0	0

Note: Kendall's rank correlation coefficients for NYCPR ($r_\tau = 0.225, p < 0.001$) and *ML-boot* ($r_\tau = 0.308, p < 0.001$).

Table 8.8 compares the two sets of grades with the Federal Accountability Status for each school. Neither of the two grades is strongly correlated with the Federal Accountability Status. But it is notable to compare the schools that are assigned As, yet have been designated for Restructuring under the Federal Accountability system. To reach this point, a school has missed AYP for at least 5 years. Of the schools in Restructuring, the NYCPR rates 38 as As—10% of all As. This highlights the lack of agreement the NYCPR has with the Federal Accountability System. Although the two scores are significantly correlated, the Kendall's rank correlation coefficient is practically-speaking negligible ($r_\tau = 0.087, p = 0.002$). By contrast, the ML-bootstrap-derived grades assign only 7 of these schools As, or less than 2%. The ML-bootstrap grades are much more strongly associated with the Federal Accountability scores ($r_\tau = 0.340, p < 0.001$).

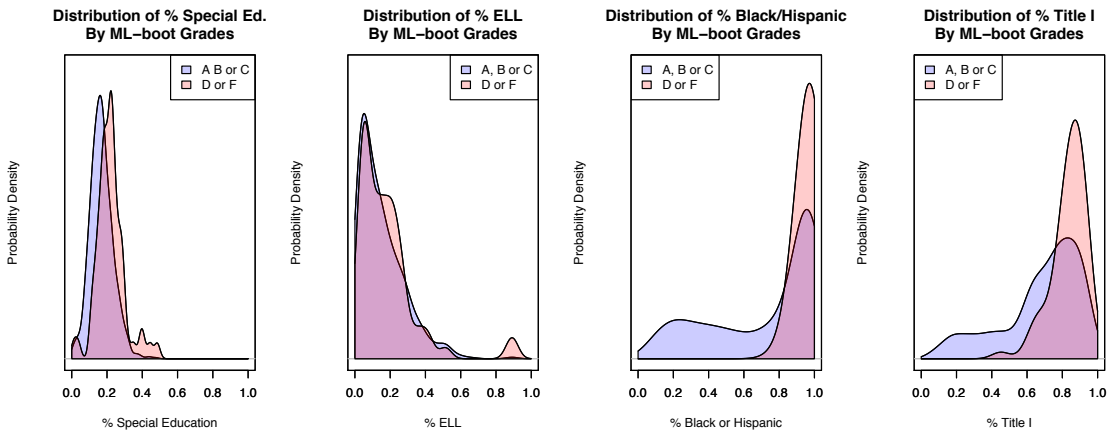

Figure 8.5**.** Comparisons of demographic distributions by ML-bootstrap estimates

A similar but less-pronounced pattern occurs with the NYC Quality Review ratings (Table 8.9). These ratings given by expert reviewers after a two-day in-person visit that included classroom observations and interviews with students, staff, and parents. They are often discussed as a qualitative counterbalance to the standardized-test-heavy focus of the Federal Accountability System. Both grading systems showed significant correlations with the Quality Review ratings, with the NYCPR ($r_\tau = 0.225, p < 0.001$) slightly lower than the *ML-boot* ($r_\tau = 0.308, p < 0.001$).

Figure 8.5 compares the distributions across the four demographics, between passing and failing schools according to the new grades based on the ML-bootstrap latent factor score estimates. In comparison with Figure 7.2, which shows the same plots based on NYCPR scores, the ML-bootstrap grades show larger differences between the distributions. In particular, the schools labeled failing here now tend to be have higher concentrations of students who receive Special Education services, who are Black or Hispanic, or who count toward Title I qualification. The equity implications of this will be discussed in the next section.

Chapter 9. Summary and Further Directions

9.1. Summary

This study examined the impact of being rated a "failing" school, here defined as receiving a "D" or "F" on the New York City Progress Report (NYCPR) in the 2006-07 school year, on school quality. There are three primary research questions investigated here:

Q1: How accurate are characterizations of school quality?

Q2: How effective are school rating systems at improving school quality?

Q3: Are there demographic differences in school quality?

To this end, a methodology for incorporating propensity score matching within a second-order growth model (PS-SGM) is proposed. Simulations indicate that this methodology exhibits improved coverage rates and reduced bias across a range of treatment effect sizes. The results of the PS-SGM estimated under a maximum likelihood regime are discussed here.

This study compares the School Quality grades under the original NYCPR formulation with those resulting from a PS-SGM. In calculating School Quality grades, the NYCPR used 15 measurements, grouped into three subdomains: School Environment, Student Performance, and Student Progress. Of these three, Student Progress was by far the most central, accounting for 60% of the final score; Student Performance coming in second, at 25%; and School Environment comprising the final 15%. These weights reflect the best judgment of stakeholders as the Progress Report system was developed over years and in response to various stakeholder feedback. It is an aspirational determination of what, collectively, stakeholders would like a

quality school to be: a school that takes students and helps them learn as much as possible, with a special focus on incoming students with the lowest starting performance.

However, this definition of presumes School Quality to be an emergent construct—completely determined by and only by the included measures—which is an untenable assumption. Switching to a latent construct perspective is more theoretically defensible (Hancock et al., 2001). The proposed PS-SGM does just that, and the results give a much different picture of School Quality. Here we see Student Progress as marginally associated with School Quality (8%). Instead, School Environment is the most associated with School Quality (57%), followed by Student Performance (35%). Compared with two other indicators of school quality, the PS-SGM scores show improved consistency with the Federal Accountability scores ($r_\tau = 0.340, p < 0.001$) and the New York City Quality Review ($r_\tau = 0.308, p < 0.001$). The NYCPR scores exhibit lower consistency with the Federal Accountability scores ($r_\tau = 0.087, p = 0.002$) and the New York City Quality Review ($r_\tau = 0.225, p < 0.001$).

Of particular note is the low correlation between the NYCPR scores and the Federal Accountability scores. One possible explanation for this is hinted at by the pattern of loadings suggested by the PS-SGM. The near-zero loading of School Progress on the overall School Quality suggests that a unidimensional definition of School Quality may be inappropriate. This is also reflected in the model fit. The factor structure that reflected the original NYCPR design exhibits extremely poor model fit (RMSEA = 0.27; TLI = 0.38). With the addition of method factors and relaxing of strict factor invariance assumptions, the model fit improves but continues to be poor (RMSEA = 0.20; TLI = 0.65). The large percentage of Heywood or ultra-Heywood cases in *ML-boot* results—much larger than would be expected based on simulation results—

adds to the evidence that a unidimensional model of School Quality is a structural misspecification (Kolenikov & Bollen, 2012).

These results support the argument that identifying failing schools by a single summary statistic is inappropriate. The ability to educate students performing at grade-level and the ability to educate students performing below grade-level may be separate dimensions. The stated goal of the NYCPR in weighting Student Progress so heavily was to "reflect each school's contribution to student academic progress, no matter where each child begins his or her journey to proficiency and beyond" (p. 2, NYCDOE). Toward this end, Student Progress was described as a way to value schools in which the incoming student body was below grade-level but by the end of the year had made significant gains. As previously discussed, the results here may indicate that this concept of a quality school may be a separate, orthogonal factor from a definition of School Quality that relies on current-year student performance. In fact, capturing this alternate dimension of quality may be exactly the aim of the NYCPR grades. By collapsing all the indicators into a unidimensional metric, however, this nuance is lost from both the measurement perspective and from the public accountability perspective.

This is of particular importance in the current regulatory environment. As Darling-Hammond notes,

> "[t]he law requires states to develop processes for identifying and supporting the lowest performing schools (the 'bottom 5 percent' of Title I schools) and those with sustained equity gaps. Although ESSA states that the set of academic measures must have greater weight than other non-academic measures in making the determination, this does not mean that a unidimensional index or grading scheme must be used as the foundation of the accountability system" (Darling-hammond et al., 2016, p. 24).

The authors continue on to offer two potential approaches to make such determinations: a weighted measures approach or the use of decision rules. If School Quality is truly a multi-dimensional construct, this would be a strong argument in favor of the use of decision rules. Even the best-designed weighted measures approach cannot accurately reflect the complexity of school quality.

The labeling of schools has important consequences. The results here estimate that schools identified for intervention by the NYCPR showed an additional increase in growth of 0.167 standard units, based on ML estimates. This is 33% of the size of the growth experienced by the control group. This suggests that, at least in this first year of the program, the receipt of a D or F was associated with a marked increase in growth. These impact sizes are consistent with other findings on the impact of failing grades on standardized test scores in the same data (Rockoff & Turner, 2010; Winters & Cowen, 2012).

The mechanism by which this change occurred is an avenue for further investigation, although it is likely it is "market accountability" (Murray & Howe, 2017) through the negative publicity and stigma associated with the failing grades, as specific consequences had not yet been attached to these first year scores (Figlio & Winicki, 2005; Rouse et al., 2013). However, principals at schools receiving a D or F were required to create written action plans for improvement, which also could have had a direct impact.

The lack of an established operational definition of School Quality makes the task of examining bias challenging. For example, a lack of differences across demographics could be due to either a true equality of outcomes; or a mis-specification of school quality that obscures differences in reality. Without a theoretical framework, it is difficult to know which is more applicable to a given situation. The approach taken here is to apply a methodological solution—

propensity score matching—that theoretically removes the confounding influence of covariates, by selecting comparable samples for comparison. The distributions of demographics in the matched control and treatment groups suggest that the matching process has accomplished this task, by averaging over multiple samples based on separate estimates of the propensity score. Thus, the estimates based on this matching process are, in principle, controlled for the influence of these demographic contexts.

The demographic patterns of the resultant grades, then, suggest a more dramatic difference in School Quality than the original NYCPR grades. This can be seen in the demographic characteristics of the median school receiving a passing grade versus a failing grade. For the original NYCPR grades, the differences between the demographic distributions for schools receiving passing grades versus those receiving failing grades are statistically significant at the $\alpha = 0.05$ level for Special Education students (median = 16.4% for passing schools; median = 18.7% for failing schools); for Black or Hispanic students (median = 89.9% for passing schools; 96.6% for failing schools); and for Title I status (median = 74.4% for passing schools; 81.1% for failing schools); but not for English Language Learners (median = 11.8% for passing schools; 11.1% for failing schools). The same pattern of statistically significant differences exists for the demographics under the PS-SGM grades. However, the differences between the median schools grow wider: for Special Education (16.2% for passing vs. 21.8% for failing); for Black or Hispanic (88.6% vs. 97.7%); and for Title I (73.3% vs. 85.8%); and for English Language Learners (11.4% vs. 14.0%). This suggests that the "educational gaps" are wider along these dimensions than is measured by the original NYCPR grades; and that the mis-specification of weights under the NYCPR may be obscuring the actual differences in school quality for students of different demographics.

9.2. Limitations

Caution should be used in over-interpreting these results. The model here is knowingly mis-specified, as the impact of Initial Quality on the treatment assignment (a failing grade) is not included in the propensity score model. However, the imputed Initial Quality distributions between the treatment and control groups were comparable (Figure 8.4). The simulation results also suggest that this mis-specification of the treatment assignment mechanism may not impact the accuracy of the treatment effect estimates. In order to explicitly include this dependence in the model, the estimation of the measurement model would have to be separated between years, in order to first impute the estimated factor scores prior to the propensity scores. This would impact the standard error estimates in unpredictable ways, potentially undermining any inferences based on these standard error estimates. In light of this, the trade-off was to allow the propensity model to be knowingly mis-specified, relying on the simulation results to assume that the treatment effect estimates are not dramatically impacted.

Indeed, the effect size estimates are consistent with those from previous studies that used regression discontinuity approaches. This study applies propensity scores to accomplish the same purpose as the regression discontinuity approach in these previous studies: to create similar groups for comparison to estimate causal effects. The consistency of the effect size estimates across multiple methodologies provides further evidence that there is in fact a positive impact on student outcomes due to school accountability systems.

Another source of concern regarding the reliability of the model are the extremely poor model fit statistics. However, the model fit statistics also suggest these results are an

improvement over the original formulation which is consistent with comparisons with the Federal Accountability ratings and Quality Review scores.

These results should not be taken as definitive evidence on the efficacy of school accountability measures. Instead, the primary message here is that adopting a latent factor perspective of School Quality can provide insight into the validity of these systems that is currently lacking. Using the statistical tools here can provide insight into the development of improved measures of School Quality.

The whole-sale adoption of a latent factor approach is not required for these findings to be useful in the development of school accountability systems. Rather, these methodologies provide insight into the nature of the underlying structures of School Quality. These insights can insert a different perspective into the conversation as a counterpoint to other sources of expert judgment. For example, the extremely low factor loadings for the Student Progress measures suggest that these indicators are unreliable measures of School Quality. Having highly unreliable indicators account for over half of a school's total rating, when these ratings have such high-stakes consequences, results in a set of ratings that may themselves be unreliable. If the goal is to reward schools that make large gains with their students, then this provides strong evidence that heavily weighting Student Progress metrics in a unidimensional measure of School Quality is not a reliable way to accomplish that goal.

There were also some constraints in the design and implementation of this study that should also be kept in mind. Although the methodology proposed here is flexible enough to incorporate many time points, the data analyzed were limited to the 2006-07 and 2007-08 school years for several reasons. One of the primary research questions is to investigate the impact of the school rating systems on school quality, and in particular the impact of receiving a poor

rating. By limiting the investigation to two years, the interactivity of effects is minimized; for example, if the range had been expanded to include a third year, receiving a low grade the second year would be a separate treatment, making the effects on the third year a mixture of the treatment in the first year and the treatment in the second year. It is not reasonable to assume that these treatments are merely linearly additive and independent. Thus, the complexity of the modeling task would increase more than exponentially with additional years.

Beyond the increase in modeling complexity, expanding the time window also creates challenges for causal inference. The underlying assumption of SUTVA, which states that the treatment of one unit does not affect the outcome of another unit, becomes untenable. Schools are not receiving their grades in isolation, but rather as part of a larger educational ecosystem. One of the consequences in publishing and disseminating school grades is that it leads to a migration of students from low-scoring schools to higher-scoring schools. This migration would suggest that a school's student population is directly impacted by not only its own grade, but the grade of other schools. By limiting this analysis to two years, this threat is minimized because the grades are published in the middle of the next school year, too late for most students to transfer schools.

9.3. Further Research

Since the initiation of this research project, the context in New York City has changed. With a change in mayoral administration, the New York City Department of Education has undergone a change in leadership. The unidimensional School Progress Report has been replaced with an informational dashboard, reporting a variety of metrics in several subdomains, without the calculation of a single School Quality rating. This transition highlights the importance of further investigation into the multi-dimensional nature of School Quality.

The type of indicators used in measuring School Quality here are limited to those used in the original NYCPR, in order to provide a more direct comparison with an existing model. However, as the current legislative regime requires states to include a larger array of indicators, latent factor methodologies—as opposed to the emergent construct perspective most often currently used—can be an invaluable in designing school accountability models with validity in mind.

Adopting a latent factor perspective provides a way to investigate how an increasingly complex set of indicators are interrelated and correspond to our ideas of what constitutes a good school. Although this study aimed to estimate the impact of an intervention in comparison to a pre-existent system, the results lay bare the insufficiency of a unidimensional definition for School Quality. There is need for more research into the basic factor structure of School Quality. The methodologies explored here can easily be expanded to investigate the potential multi-dimensional nature of School Quality. A second-order latent growth model is especially appropriate for this, as not only can quality be modeled across years, but also assumptions regarding the static or evolving character of School Quality could be tested.

In addition, the incorporation of student-level data and explicit modeling of individual student growth over time in a multi-level model would obviate the need to rely on aggregate measures derived from the same sources. Current systems all include student current-year performance and student growth as two separate sets of indicators, which are aggregated at the school level. If instead the data were modeled at the student level, changes in performance and growth could be modeled as a single quadratic latent factor growth model.

On the technical side, there are also several avenues for further development. Other propensity score methodologies—such as stratification or weighting—could also be applied,

especially to estimate treatment effects at different points. Also, as the simulation here only explored a limited set of simulation parameters, the performance of such models still needs be explored with a larger array of covariates, different misspecifications of propensity score models, direct covariate effects on outcomes, or varying sample sizes for example.

There are also avenues for investigation in the areas of Bayesian prior setting. One of the advantages of a Bayesian approach is the ability to incorporate expert knowledge through the appropriate prior definitions. In the absence of strong expert knowledge, however, often default non-informative priors are used. The impact of different non-informative prior specifications on the resulting model parameter estimates and model convergence is an important area of exploration.

Another area of development that could be particularly useful in the policy context is the development of robust estimate errors for school factor scores. Within the Bayesian regime, these standard errors would be a natural result of the posterior distribution. However, as was the case in this study, computational resources may be a limiting factor in utilizing this approach for Bayesian factor scores.

The methodology proposed here is also broadly applicable to a variety of other contexts. The PS-SGM model can be applied to any situation where there is an intervention that aims to effect change in an unobservable latent construct. These types of situations arise often in educational or psychological contexts. From individual student learning or teacher quality to neuroticism or self-esteem, many educational or psychological theories are built on models for constructs that cannot be directly observed. This approach then provides a powerful tool to investigate the efficacy of a broad array of interventions.

Bibliography

Abadie, A., & Imbens, G. W. (2006). Large Sample Properties of Matching Estimators For Average Treatment Effects. *Econometrica*, *74*(1), 235–267. http://doi.org/10.1111/j.1468-0262.2006.00655.x

Abadie, A., & Imbens, G. W. (2008). On the Failure of the Bootstrap for Matching Estimators. *Econometrica*, *76*(6), 1537–1557. http://doi.org/10.3982/ECTA6474

Alvarez, R. M., & Levin, I. (2014). Uncertain Neighbors : Bayesian Propensity Score Matching for Causal Inference.

An, W. (2010). Bayesian Propensity Score Estimators: Incorporating Uncertainties in Propensity Scores Into Causal Inference. *Sociological Methodology*, *40*(1), 151–189.

Angrist, J., & Lavy, V. (2009). The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial. *The American Economic Review*, *99*(4), 1384–1414.

Bishop, J., Geiser, C., & Cole, D. A. (2015). Modeling Latent Growth With Multiple Indicators: A Comparison of Three Approaches. *Psychological Methods*, *20*(1), 43–62. http://doi.org/10.1037/met0000018

Bokhari, F. a S., & Schneider, H. (2011). School accountability laws and the consumption of psychostimulants. *Journal of Health Economics*, *30*(2), 355–72. http://doi.org/10.1016/j.jhealeco.2011.01.007

Bollen, K. A. (2002). Latent Variables in Psychology and the Social Sciences. *Annual Review of Psychology*, *53*, 605–634. http://doi.org/10.1146/annurev.psych.53.100901.135239

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective* (Vol 467). John Wiley & Sons.

Bollen, K. A., & Lennox, R. (1991). Conventional Wisdom on Measurement: A Structural Equation Perspective. *Psychological Bulletin*, *110*(2), 305–314. http://doi.org/10.1037/0033-2909.110.2.305

Booher-Jennings, J. (2005). Below the Bubble: "Educational Triage" and the Texas Accountability System. *American Educational Research Journal*, *42*(2), 231–268. http://doi.org/10.3102/00028312042002231

Carnoy, M., & Loeb, S. (2002). Does External Accountability Affect Student Outcomes? A Cross-State Analysis. *Educational Evaluation and Policy Analysis*, *24*(4), 305–331. http://doi.org/10.3102/01623737024004305

Chakrabarti, R., & Schwartz, N. (2013). *Unintended Consequences of School Accountability Policies : Evidence from Florida and Implications for New York*.

Chan, W., & Bentler, P. M. (1998). Covariance structure analysis of ordinal ipsative data. *Psychometrika*, *63*(4), 369–399. http://doi.org/10.1007/BF02294861

Chay, K. Y., McEwan, P. J., & Urquiola, M. (2005). The Central Role of Noise in Evaluating Interventions That Use Test Scores to Rank Schools. *American Economic Review*, *95*(4), 1237–1258.

Chen, F. F., Sousa, K. H., & West, S. G. (2009). Teacher's Corner: Testing Measurement Invariance of Second-Order Factor Models. *Structural Equation Modeling*, *12*(3), 368–390. http://doi.org/10.1207/s15328007sem1203_7

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics*, *93*(9–10), 1045–1057. http://doi.org/10.1016/j.jpubeco.2009.06.002

Coe, R. (2009). Unobserved but not unimportant: the effects of unmeasured variables on causal attributions. *Effective Education*, *1*(2), 101–122. http://doi.org/10.1080/19415530903522519

Cohen, P., Cohen, J., Teresi, J., Marchi, M., & Velez, C. N. (1990). Problems in the Measurement of Latent Variables in Structural Equations Causal Models. *Applied Psychological Measurement*, *14*(2), 183–196. http://doi.org/10.1177/014662169001400207

Cox, D. R. (1958). *Planning of Experiments*. New York: Wiley.

Darling-hammond, L., Bae, S., Cook-Harvey, C. M., Lam, L., Mercer, C., Podolsky, A., & Stosich, E. L. (2016). *Pathways to New Accountability Through the Every Student Succeeds Act*. Palo Alto.

Darling-Hammond, L. (2007). Race, inequality and educational accountability: the irony of 'No Child Left Behind.' *Race Ethnicity and Education*, *10*(3), 245–260. http://doi.org/10.1080/13613320701503207

Dee, T. S., & Jacob, B. (2011). The Impact of No Child Left Behind on Student Achievement. *Journal of Policy Analysis and Management*, *30*(3), 418–446. http://doi.org/10.1002/pam

Denning, P. J. (1983). A nation at risk: the imperative for educational reform. *The Elementary School Journal*, *84*(2), 112–130. http://doi.org/10.1145/358150.358154

Denwood, M. J. (2016). runjags: An R Package Providing Interface Utilities, Model Templates, Parallel Computing Methods and Additional Distributions for MCMC Models in JAGS. *Journal of Statistical Software*, *71*(9), 1–25.

Dobbie, W., & Fryer, R. G. (2013a). Getting Beneath the Veil of Effective Schools: Evidence From New York City. *American Economic Journal: Applied Economics*, *5*(4), 28–60.

Dobbie, W., & Fryer, R. G. (2013b). The Medium-Term Impacts of High-Achieving Charter Schools on Non-Test Score Outcomes .

Duncan, T. E., & Duncan, S. C. (2009). The ABC's of LGM: An Introductory Guide to Latent Variable Growth Curve Modeling. *Social and Personality Psychology Compass*, *3*(6), 979–991. http://doi.org/10.1111/j.1751-9004.2009.00224.x.The

Engberg, J., Gill, B., Zamarro, G., & Zimmer, R. (2012). Closing schools in a shrinking district: Do student outcomes depend on which schools are closed? *Journal of Urban Economics*, *71*(2), 189–203. http://doi.org/10.1016/j.jue.2011.10.001

Faubert, V. (2009). *School Evaluation: Current Practices in OECD Countries and a Literature Review* (OECD Education Working Papers No. 42).

Ferrer, E., Balluerka, N., & Widaman, K. F. (2008). Factorial Invariance and the Specification of Second-Order Latent Growth Models. *Methodology*, *4*(1), 22–36. http://doi.org/10.1027/1614-2241.4.1.22.Factorial

Figlio, D. N., & Kenny, L. W. (2009). Public sector performance measurement and stakeholder support. *Journal of Public Economics*, *93*(9–10), 1069–1077. http://doi.org/10.1016/j.jpubeco.2009.07.003

Figlio, D. N., & Lucas, M. E. (2004). What's in a Grade? School Report Cards and the Housing Market. *The American Economic Review*, *94*(3), 591–604.

Figlio, D. N., & Rouse, C. E. (2006). Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics*, *90*(1–2), 239–255. http://doi.org/10.1016/j.jpubeco.2005.08.005

Figlio, D. N., & Winicki, J. (2005). Food for thought: the effects of school accountability plans on school nutrition. *Journal of Public Economics*, *89*(2–3), 381–394. http://doi.org/10.1016/j.jpubeco.2003.10.007

Geiser, C., Keller, B. T., & Lockhart, G. (2013). First- Versus Second-Order Latent Growth Curve Models: Some Insights From Latent State-Trait Theory. *Structural Equation Modeling*, *20*(3), 479–503. http://doi.org/10.1080/10705511.2013.797832

Gelman, A., & Hill, J. (2007). *Data analysis using regression and hierarchical/multilevel models*.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S. W., & Whitehurst, G. J. (2010). *Evaluating Teachers: The Important Role of Value-Added. Brookings Institution*. Retrieved from http://www.eric.ed.gov/ERICWebPortal/recordDetail?accno=ED512829

Gootman, E., & Medina, J. (2007, November 6). 50 New York Schools Fail Under Rating System. *The New York Times*. Retrieved from https://www.nytimes.com/2007/11/06/education/06reportcards.html

Grissmer, D., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving Student Achievement: What State NAEP Test Scores Tell Us*.

Gyurko, J., & Henig, J. R. (2010). Strong vision, learning by doing, or the politics of muddling through: New York City. In *Between Public and Private: Politics, Governance, and the New Portfolio Models for Urban School Reform* (pp. 91–126). Cambridge, MA: Harvard Education Press.

Hahn, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, *66*(2), 315–331.

Hamilton, L., Berends, M., & Stecher, B. (2005). *Teachers' responses to standards-based accountability*. Retrieved from http://192.5.14.43/content/dam/rand/pubs/working_papers/2005/RAND_WR259.pdf

Hancock, G. R., Kuo, W., & Lawrence, F. R. (2001). An Illustration of Second-Order Latent Growth Models. *Structural Equation Modeling*, *8*(3), 470–489. http://doi.org/10.1207/S15328007SEM0803

Hansen, B. B. (2004). Full Matching in an Observational Study of Coaching for the SAT. *Journal of the American Statistical Association*, *99*(467), 609–618. http://doi.org/10.1198/016214504000000647

Hansen, B. B., & Bowers, J. (2008). Covariate Balance in Simple, Stratified and Clustered Comparative Studies. *Statistical Science*, *23*(2), 219–236. http://doi.org/10.1214/08-STS254

Hansen, B. B., & Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, *15*(3), 609–627. http://doi.org/10.1198/106186006X137047

Hanushek, E. a., Kain, J. F., & Rivkin, S. G. (2004). Disruption versus Tiebout improvement: the costs and benefits of switching schools. *Journal of Public Economics*, *88*(9–10), 1721–1746. http://doi.org/10.1016/S0047-2727(03)00063-X

Hanushek, E. a., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, *24*(2), 297–327. http://doi.org/10.1002/pam.20091

Hastings, J. S., & Weinstein, J. M. (2008). Information, school choice, and academic achievement: evidence from two experiments. *The Quarterly Journal of Economics*, *123*(4), 1373–1414.

Hayduk, L. A., & Littvay, L. (2012). Should researchers use single indicators, best indicators, or multiple indicators in structural equation models? *BMC Medical Research Methodology*, *12*. http://doi.org/10.1186/1471-2288-12-159

Heckman, J. J., & Hotz, V. J. (1989). Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs : The Case of Manpower Training. *Journal of the American Statistical Association*, *84*(408), 862–874.

Hill, P. T. (2011). Leadership and governance in New York City school reform. In J. A. O'Day,

C. S. Bitter, & L. M. Gomez (Eds.), *Education reform in New York City: Ambitious change in the nation's most complex school system* (pp. 17–32). Cambridge, MA: Harvard University Press.

Hirano, K., & Imbens, G. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, *2*(3), 259–278. Retrieved from http://dx.doi.org/10.1023/A:1020371312283

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica*, *71*(4), 1161–1189.

Hooge, E. (2012). Making multiple school accountability work (pp. 93–113).

Hooge, E., Burns, T., & Wilkoszewski, H. (2012). *Looking Beyond the Numbers : Stakeholders and Multiple School Accountability* (OECD Education Working Papers No. 85).

Hoshino, T. (2007). Doubly robust-type estimation for covariate adjustment in latent variable modeling. *Psychometrika*, *72*(4), 535–549. Retrieved from http://www.springerlink.com/index/10.1007/s11336-007-9007-2

Hoshino, T. (2008). A Bayesian propensity score adjustment for latent variable modeling and MCMC algorithm. *Computational Statistics & Data Analysis*, *52*(3), 1413–1429. http://doi.org/10.1016/j.csda.2007.03.024

Hoshino, T., Kurata, H., & Shigemasu, K. (2006). A Propensity Score Adjustment for Multiple Group Structural Equation Modeling. *Psychometrika*, *71*(4), 691–712. Retrieved from http://link.springer.com/article/10.1007/s11336-005-1370-2

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimantalists and observationalists about causal inference. *Journal of the Royal Statistical Society*, *171*(2), 481–502. http://doi.org/10.1111/j.1467-985X.2007.00527.x

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

Jacob, B. A. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, *89*(5–6), 761–796. http://doi.org/10.1016/j.jpubeco.2004.08.004

Jacob, B. A., & Levitt, S. D. (2003). Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *The Quarterly Journal of Economics*, *118*(3), 843–878.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Scientific Software International.

Jorgensen, T. D., Pornprasertmanit, S., Miller, P., & Schoemann, A. (2017). simsem: SIMulated

Structural Equation Modeling.

Kane, T. J., & Staiger, D. O. (2002). The Promise and Pitfalls of Using Imprecise School Accountability Measures. *Journal of Economic Perspectives*, *16*(4), 91–114. http://doi.org/10.1257/089533002320950993

Kaplan, D. (1999). An Extension of the Propensity Score Adjustment Method for the Analysis of Group Differences in MIMIC Models. *Multivariate Behavioral Research*, *34*(4), 467. http://doi.org/10.1207/S15327906MBR3404

Kaplan, D. (2012). Bayesian Structural Equation Modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 650–673). The Guilford Press.

Kaplan, D., & Chen, J. (2012). A Two-step Bayesian Approach For Propensity Score Analysis: Simulations and Case Study. *Psychometrika*, *77*(3), 581–609.

Kaplan, D., & Chen, J. (2014). Bayesian Model Averaging for Propensity Score Analysis. *Multivariate Behavioral Research*, *49*(6), 505–517. http://doi.org/10.1080/00273171.2014.928492

Kim, J. S., & Sunderman, G. L. (2005). Measuring Academic Proficiency Under the No Child Left Behind Act: Implications for Educational Equity. *Educational Researcher*, *34*(8), 3–13. http://doi.org/10.3102/0013189X034008003

King, G., & Zeng, L. (2006). The dangers of extreme counterfactuals. *Political Analysis*, *14*(2), 131–159. http://doi.org/10.1093/pan/mpj004

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). What Do Test Scores in Texas Tell Us? *Education Policy Analysis Archives*, *8*(49).

Koedel, C., & Betts, J. (2007). *Re-Examining the Role of Teacher Quality In the Educational Production Function*. Retrieved from http://econ.missouri.edu/working-papers/2007/wp0708_koedel.pdf

Kolenikov, S., & Bollen, K. A. (2012). *Testing Negative Error Variances: Is a Heywood Case a Symptom of Misspecification? Sociological Methods and Research* (Vol. 40). http://doi.org/10.1177/0049124112442138

Kolodner, M. (2010, July 20). State test scores improved because of easier tests, not more learning: study. *The New York Daily News*. Retrieved from http://www.nydailynews.com/new-york/education/state-test-scores-improved-easier-tests-not-learning-study-article-1.467016

Koretz, D. (2009). *Measuring up: what educational testing really tells us*. Harvard University Press.

Kress, S., Zechmann, S., & Schmitten, J. M. (2011). When Performance Matters : Consequential Accountability in Public Education. *Harvard Journal On Legislation*, *48*(1), 185–234.

Ladd, H. F., & Zelli, A. (2002). School-Based Accountability in North Carolina: The Responses of School Principals. *Educational Administration Quarterly*, *38*(4), 494–529. http://doi.org/10.1177/001316102237670

Lee, S. (2007). *Structural Equation Modeling: A Bayesian Approach*.

Leite, W. L. (2007). A comparison of latent growth models for constructs measured by multiple items. *Structural Equation Modeling*, *14*(4), 581–610. http://doi.org/10.1080/10705510701575438

Leite, W. L., Sandbach, R., Jin, R., MacInnes, J. W., & Jackman, M. G.-A. (2012). An Evaluation of Latent Growth Models for Propensity Score Matched Groups. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(January 2015), 437–456. http://doi.org/10.1080/10705511.2012.687666

Linn, R. L., & Haug, C. (2002). Stability of School-Building Accountability Scores and Gains. *Educational Evaluation and Policy Analysis*, *24*(1), 29–36. http://doi.org/10.3102/01623737024001029

Martin, A. D., Quinn, K. M., & Park, J. H. (2011). MCMCpack : Markov Chain Monte Carlo in R. *Journal of Statistical Software*, *42*(9), 1–21. http://doi.org/10.18637/jss.v042.i09

McArdle, J. J. (1988). Dynamic but Structural Equation Modeling of Repeated Measures Data. In N. J.R. & C. R.B. (Eds.), *Handbook of Multivariate Experimental Psychology. Perspectives on Individual Differences*. Boston, MA: Springer.

McArdle, J. J., & Epstein, D. (1987). Latent Growth Curves within Developmental Structural Equation Models. *Child Development*, *58*(1), 110–133.

McCandless, L. C., Douglas, I. J., Evans, S. J., & Smeeth, L. (2010). Cutting Feedback in Bayesian Regression Adjustment for the Propensity Score. *The International Journal of Biostatistics*, *6*(2). http://doi.org/10.2202/1557-4679.1205

McCandless, L. C., Gustafson, P., & Austin, P. C. (2009). Bayesian propensity score analysis for observational data. *Statistics in Medicine*, *28*, 94–112. http://doi.org/10.1002/sim

McCandless, L. C., Richardson, S., & Best, N. (2012). Adjustment for Missing Confounders Using External Validation Data and Propensity Scores. *Journal of the American Statistical Association*, *107*(497), 40–51. http://doi.org/10.1080/01621459.2011.643739

Meredith, W., & Horn, J. (2001). The role of factorial invariance in modeling growth and change. In L. M. Collins & A. G. Sayer (Eds.), *Decade of behavior. New methods for the analysis of change* (pp. 203–240). Washington, DC, US: American Psychological Association.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*(1), 107–122. http://doi.org/10.1007/BF02294746

Merkle, E. C., & Rosseel, Y. (2015). blavaan: Bayesian structural equation models via parameter expansion. *ArXiv*, (Rosseel 2012). Retrieved from http://arxiv.org/abs/1511.05604

Mintrop, H., & Sunderman, G. L. (2009). Predictable Failure of Federal Sanctions-Driven Accountability for School Improvement--And Why We May Retain It Anyway. *Educational Researcher*, *38*(5), 353–364. http://doi.org/10.3102/0013189X09339055

Murray, K., & Howe, K. R. (2017). education policy analysis archives Neglecting Democracy in Education Policy : A-F School Report Card Accountability Systems.

Muthen, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557–585. http://doi.org/10.1007/BF02296397

Muthén, B., & Asparouhov, T. (2011). Bayesian SEM : A more flexible representation of substantive theory Web tables. *Psychological Methods*, *17*(3), 313–335.

New York City Department of Education. (2007). *Educator Guide: The New York City Progress Report*. New York.

Oort, F. J. (2001). Three-mode models for multivariate longitudinal data. *British Journal of Mathematical and Statistical Psychology*, *54*(1), 49–78. http://doi.org/10.1348/000711001159429

Papay, J. P., Murnane, R. J., & Willett, J. B. (2010). The Consequences of High School Exit Examinations for Low-Performing Urban Students: Evidence From Massachusetts. *Educational Evaluation and Policy Analysis*, *32*(1), 5–23. http://doi.org/10.3102/0162373709352530

Pearl, J. (2000). *Causality*. New York: Cambridge University Press.

Peck, C. (2014). Paradigms, Power, and PR in New York City: Assessing Two School Accountability Implementation Efforts. *Education Policy Analysis Archives*, *22*(114). http://doi.org/10.1080/03057260903142269

Plummer, M. (2015). JAGS Version 4.0 user manual, (August), 0–41.

Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology (Cambridge, Mass.)*, *11*(5), 550–560. http://doi.org/10.1097/00001648-200009000-00011

Rockoff, B. J., & Turner, L. J. (2010). Short-Run Impacts of Accountability on School Quality. *American Economic Journal: Economic Policy*, *2*(4), 119–147.

Rosenbaum, P. R. (1986). Dropping out of High School in the United States: An Observational Study. *Journal of Educational and Behavioral Statistics*, *11*(3), 207–224. http://doi.org/10.3102/10769986011003207

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American*

Statistical Association, *84*(408), 1024–1032. http://doi.org/10.1080/01621459.1989.10478868

Rosenbaum, P. R., & Rubin, D. B. (1983a). Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome. *Journal of the Royal Statistical Society*, *45*(2), 212–218.

Rosenbaum, P. R., & Rubin, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Rothstein, J. (2009). Student Sorting and Bias In Value-Added Estimation: Selection On Observables and Unobservables. *Education Finance and Policy*, *4*(4), 537–571.

Rouse, C. E., Hannaway, J., Goldhaber, D., & Figlio, D. (2013). Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure. *American Economic Journal: Economic Policy*, *5*(2), 251–281. http://doi.org/10.1257/pol.5.2.251

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. http://doi.org/10.1037/h0037350

Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomizaion. *The Annals of Statistics*, *6*(1), 34–58.

Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics*, *12*(4), 1151–1172. http://doi.org/10.1214/aos/1176348654

Rubin, D. B. (2008). For Objective Causal Inference, Design Trumps Analysis. *The Annals of Applied Statistics*, *2*(3), 808–840.

Saarela, O., Stephens, D. A., Moodie, E. E. M., & Klein, M. B. (2015). On Bayesian estimation of marginal structural models. *Biometrics*, *71*(2), 279–288. http://doi.org/10.1111/biom.12269

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, *64*(1), 37–52. http://doi.org/10.1007/BF02294318

Song, X.-Y., & Lee, S.-Y. (2006). Maximum Likelihood Approach for Multisample Nonlinear Structural Equation Models With Missing Continuous and Dichotomous Data. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(3), 325–351. http://doi.org/10.1207/s15328007sem1303

Song, X.-Y., Lee, S.-Y., & Hser, Y.-I. (2008). A two-level structural equation model approach for analyzing multivariate longitudinal responses. *Statistics In Medicine*, *27*, 3017–3041.

http://doi.org/10.1002/sim

Splawa-Neyman, J., Dabrowska, D. M., & Speed, T. P. (1990). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, *5*(4), 465–472. http://doi.org/10.1214/ss/1177012031

Springer, M. G. (2008). The influence of an NCLB accountability plan on the distribution of student test score gains. *Economics of Education Review*, *27*(5), 556–563. http://doi.org/10.1016/j.econedurev.2007.06.004

Stecher, B. M., Barron, S. L., Chun, T., & Ross, K. (2000). *The Effects of the Washington State Education Reform on Schools and Classrooms*.

Stoel, R., van den Wittenboer, G., & Hox, J. (2004). Methodological Issues in the Application of the Latent Growth Curve Model. In K. van Montfort, J. Oud, & A. Satorra (Eds.), *Recent developments on Structural Equation Models: Theory and Applications* (pp. 241–262). http://doi.org/10.1007/978-1-4020-1958-6

Thoemmes, F. J., & Kim, E. S. (2011). A Systematic Review of Propensity Score Methods in the Social Sciences. *Multivariate Behavioral Research*, *46*(1), 90–118. http://doi.org/10.1080/00273171.2011.540475

Tu, W., & Zhou, X.-H. (2002). A Bootstrap Confidence Interval Procedure for the Treatment Effect Using Propensity Score Subclassification. *Health Services & Outcomes Research Methodology*, *3*, 135–147.

Vanderweele, T. J. (2008). Sensitivity analysis: distributional assumptions and confounding assumptions. *Biometrics*, *64*(2), 645–9. http://doi.org/10.1111/j.1541-0420.2008.01024.x

von Oertzen, T., Hertzog, C., Lindenberger, U., & Ghisletta, P. (2010). The effect of multiple indicators on the power to detect inter-individual differences in change. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 627–646. http://doi.org/10.1348/000711010X486633

Whitesell, E. (2015). Do you see what I see? The impact of school accountability on parent, teacher, and student perceptions of the school environment. *Paper Presented at the Association for Education Finance and Policy Annual Conference*.

Winters, M. a., & Cowen, J. M. (2012). Grading New York: Accountability and Student Proficiency in America's Largest School District. *Educational Evaluation and Policy Analysis*, *34*(3), 313–327. http://doi.org/10.3102/0162373712440039

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian Mediation Analysis. *Psychological Methods*, *14*(4), 301–322. http://doi.org/10.1037/a0016972

Yung, Y.-F., Thissen, D., & McLeod, L. D. (1999). On the Relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, *64*(2), 113–128. Retrieved from http://link.springer.com/article/10.1007/BF02294531

Zhao, S., van Dyk, D. A., & Imai, K. (2013). Propensity-Score Based Methods for Causal Inference in Observational Studies with Fixed Non-Binary Treatments.

Zigler, C. M. (2016). The Central Role of Bayes Theorem for Joint Estimation of Causal Effects and Propensity Scores. *The American Statistician*, *70*(1), 47--54. http://doi.org/10.1080/00031305.2015.1111260

Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., & Dominici, F. (2013). Model Feedback in Bayesian Propensity Score Estimation. *Biometrics*, *69*(1), 263–273. http://doi.org/10.1111/j.1541-0420.2012.01830.x