A NATIONAL STUDY COMPARING CHARTER AND TRADITIONAL PUBLIC SCHOOLS USING PROPENSITY SCORE ANALYSIS

by

Jason M. Bryer

A Dissertation Submitted to the
University at Albany, State University of New York
In Partial Fulfillment of
the Requirements for the Degree of
Doctor of Philosophy

School of Education
Department of Educational and Counseling Psychology
Division of Educational Psychology & Methodology
2014

ABSTRACT

The concept of school choice within the United States is not new. Private schools have been educating students since the founding of the United States. However, in 1988, Ray Budde proposed an alternative approach to school choice that has come to be known as charter schools (Kolderie, 2005). Unlike their private school counterparts, charter schools receive public funding, but they are relieved of many of the bureaucratic and regulatory constraints public schools adhere to, but are still held accountable for student performance. Despite claims by charter school advocates that charter schools are performing as well if not better than the public school counterparts (see e.g. Allen, Consolettie, & Kerwin, 2009), studies provide mixed results with regard to charter school performance (see e.g. Braun, Jenkins, & Grigg, 2006a; Center for Research on Education Outcomes, 2009; Hubbard & Kulkarni, 2009). Ultimately, there is agreement that more research is necessary to address the question of whether charter schools provide substantially better academic experiences for students.

This study includes development of new methods designed for observational data analysis to investigate the question of whether students who attend charter schools outperform their public school counterparts on two key academic domains: reading and mathematics. The new methods represent extensions of modern methods for propensity score analysis and aim to reduce if not eliminate selection bias in the context of clustered data. Charter schools are, by definition, schools of choice, and this means that observational data methods are preferred for comparing such schools with others. In observational data contexts, simple comparisons of two groups such as traditional public and charter schools cannot help but ignore the inherent and systematic differences between the two groups. However, given well-designed observational studies and appropriate analysis methods, the effects of the selection bias can be reduced, if not eliminated. The end result is that the usual simple comparisons of two independent groups are replaced by comparisons that

make adjustments for covariate differences.

This is done utilizing a class of statistical procedures introduced by Rosenbaum and Rubin (1983) called propensity score analysis. Propensity score analysis has seen considerable increased use in the social sciences within the last few years (Arpino & Mealli, 2008). However, its use in situations where multilevel, or clustered data are of interest have been limited (Thoemmes & Kim, 2011). Using data from the 2009 National Assessment of Educational Progress (NAEP) for mathematics and reading at grades four and eight, estimates of the differences between charter and public schools will be calculated at two levels, namely at the state and national levels. Given the variability of charter schools laws across states, it is important to consider the impact of clustering. Analyses will be conducted using the newly developed `multilevelPSA` package (Bryer, 2011) in R (R Development Core Team, 2008). Specifically, propensity scores will be estimated within each state and these will be used for matching or stratification of students within each state. Comparisons of specific students, or groups of students, will in all cases be done within states. Effects will then be aggregated to provide state and national effect estimates.

As with all propensity score analyses, it is preferable to utilize multiple methods for estimating propensity scores (see e.g. Stuart, 2010; Rosenbaum, 2012). Doing so can help to provide confidence that results reflect what the data have to say, and is not merely an artifact of model specification or method choice. This study will utilize three overall approaches to propensity score analysis, namely stratification, matching, and multilevel stratification. Lastly, the use of graphics will be employed to evaluate balance and outcome differences using methods (functions) found in Helmreich and Pruzek (2009).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Since the opening of the first charter school in Minnesota in 1991, the United States[1] has increasingly embraced charter schools as an important option for educational reform. In the last 10 years alone, the number of charter schools has grown from 507 in the 1998-1999 school year to 4,561 in the 2007-2008 school year (see figure 1; Center for Education Reform, 2010). Currently, 40 states and the District of Columbia have charter school laws (see appendix A for enrollment by state & appendix B for a thematic map of the U.S. depicting the number of operating charter schools as of 2008). And, given Arne Duncan's appointment as Secretary of Education by President Barack Obama and the *Race to the Top* program, charter school growth and support is unlikely to slow in the near future.

In principal, charter schools have opted out of bureaucratic rules and union contracts in exchange for gaining academic autonomy in exchange for accountability and better academic environments for students (Wells, 2002). The idea is that, under this framework, teachers, administrators, students, and the community that comprise the charter school would be free to innovate. It is also the assumption that charter schools would serve as experimental schools where the innovations would inform reform of public education at large. However, some supporters argue for the eventual replacement of traditional public schools with charter schools (c.f. Ravitch, 2013), as further exemplified by the attempted school voucher legislation during the second Bush Administration and required increased cap on the number of charter schools within states as a requirement of the *Race to the Top* initiative of the Obama Administration..

Clearly charter schools have become a popular vehicle for educational reform among parents as well. The Center for Education Reform (2008) reports that 59% of charter schools have waiting lists averaging 198 students. Charter schools provide

[1]Though this study focuses on charter schools in the U.S., Canada (Foundations for the Future Charter Academy, 2007), Chile (Larrañaga, 2004), England (?, ?), Germany (Herbst, 2006), and New Zealand (Lander, 2001) also have charter schools.

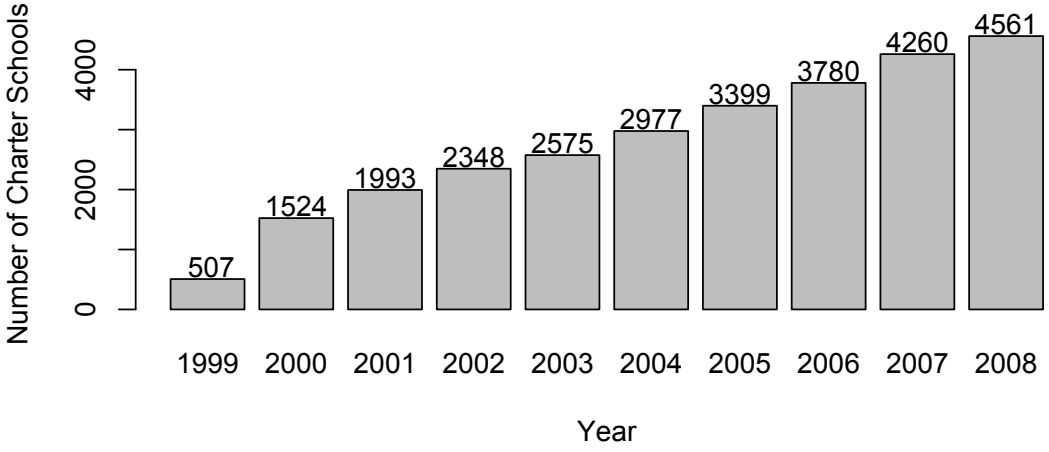


Figure 1: Charter School Growth 1999-2008

an apparent choice to parents and are copacetic to the United State's individualistic culture (see e.g., Hofstede & Hofstede, 2004; Maccall, 1847; Swart, 1962). Moreover, like so many other fields, school reform has further emphasized marketization and privatization (Wells, 2002). The influence of capitalism on education is not new. A major contributor to the expanded role of education during the industrial revolution is capitalism itself. That is, education expanded its initial purpose of providing a minimally informed electorate to providing an educated work force, not to mention keeping children off the streets as child labor laws came into existence. However, the shift of capitalistic principles from being the inspiration of educational reform to being the educational reform has profound implications.

Proponents of charter schools argue that public schools have been bogged down by bureaucracy and union contracts. Freeing schools of these requirements then allows teachers and schools to innovate, which in theory leads to increased student performance. The principled argument is the "market metaphor" (Wells, 2002). That is, if schools were forced to compete for "customers" (i.e. students), then the differentiating factor between schools would be the quality of education.

Opponents on the other hand have questioned the accountability, equity, effectiveness, and sustainability of charter schools. Several studies have shown that charter schools are not only failing to increase student performance, in many instances they are performing well under their traditional public school counterparts (see e.g., Center for Research on Education Outcomes, 2009; ?, ?; Nelson, Rosenberg, & Meter, 2004). Still, others argue whether charter schools may be a solution in search of a problem. Carnoy, Jacobsen, Mishel, and Rothstein (2005) in summarizing the controversy that ensued after the Nelson et al. (2004) study argue that:

> If, however, charter schools are not improving the achievement of disadvantages children, it may be that the cause of low student performance is not bureaucratic rules but something else. When a treatment is based on a diagnosis, and the treatment doesn't work, it is prudent to examine not only whether the treatment should be improved, but also whether the diagnosis might be flawed. (Carnoy et al., 2005)

Issues with Charter School Research

The issues surrounding charter schools are large in scope. However, given the implications for the current and future generations of students, the question must be explored using the best data and methods available. As Betts and Hill (2006) point out, there are three major obstacles to addressing the question of "whether students in charter schools are learning more or less than they would have learned in conventional schools" (p. 1), namely:

1. The issues of counterfactuals. That is, there are several barriers to determine the causal relationship between school choice and learning.

2. The variation in types of charter schools.

3. The nature of student achievement. Research has shown there are many other factors that contribute to student success including, but not limited to, socioeconomic status, parents education, motivation, etc. The ability to

decipher how school choice contributes to student learning in the context of all the other factors proves difficult.

Though these issues are significant, they can be to a large extent reasonably addressed. We will not claim to fully account for these issues, however given the need for evidence to inform policy makers regarding charter school effectiveness, we will attempt to address these issues using the best data and methods available while clearly stating the limitations.

Issue one will be dealt with in more detail in chapter three. However, in short, the propensity score analysis (PSA) proposed for this study is arguably, assuming proper implementation, the best approach to estimating causal inferences short of well designed randomized experiments. Of course in the context of an observational study the fundamental problem of causal inference Holland (1986) remains, but limitations of this will be addressed.

The issue of charter school variation is often cited in critiques of national or large scale charter school studies. Given that the charter school debate is a national debate with implications at the Federal level as exemplified by the *No Child Left Behind* legislation of the George W. Bush Administration and the *Race to the Top* policy of the Barak Obama Adminsitration, large scales studies are not only necessary, they are critical. If charter schools are to wholly be offered as an alternative to traditional public schools, then charter schools as a whole must be evaluated against public schools as a whole. More specifically, we wish not to evaluate whether a particular charter school, or type of charter school, is better, but whether the entire charter school concept is a better approach for educational reform.

Lastly, the environmental, social, community, and cultural factors that contribute to a student's academic achievement are often significantly underestimated. Often educational reform, as exemplified by the *No Child Left Behind* Act and *Race to the Top*, places the responsibility solely on the school

without consideration of the context in which the school operates. We are encouraged by President Obama's remarks to his first Joint Session of Congress (?, ?):

> These education policies will open the doors of opportunity for our children. But it is up to us to ensure they walk through them. In the end, there is no program or policy that can substitute for a mother or father who will attend those parent/teacher conferences, or help with homework after dinner, or turn off the TV, put away the video games, and read to their child. I speak to you not just as a President, but as a father when I say that *responsibility for our children's education must begin at home* [emphasis added].

Though we must be acutely aware and acknowledge the fact that schools are merely one factor of many that contribute to a students academic achievement, it does not preclude us from evaluating schools for their part. Similar to issue one, we argue that PSA provides an approach to best approximate the effects of school choice.

Research Questions & Objectives

The primary focus of this study is the development of a new set of methods for propensity score analysis with multilevel, or clustered, data. One of these aims is to show how graphics can be used to address research questions in the context of multilevel propensity score analysis; another is to describe and illustrate the key features of a new package of R functions to facilitate multilevel propensity score analyses, vis-à-vis the `multilevelPSA` package in R. Moreover, these new multilevel methods for propensity score analysis will be presented within the context of more traditional methods for propensity score analysis, namely stratification and matching. Not only will this show how these new methods perform with regard to more established methods, they may show with the use of modern graphics, how clusters may vary.

The newly developed `multilevelPSA` package will be shown to provide an effective means of estimating and visualizing propensity score results with clustered (multilevel) data (these procedures are discussed more fully below). Moreover, the use of pre-existing visualization procedures such as loess regression plots, density

plots, as well as the PSA balance and assessment plots introduced by Helmreich and Pruzek (2009), can provide critical insight into the analysis and eventual interpretation of results. More succinctly, the presentation of graphics in this study are not merely provided for diagnostic or descriptive purposes, but are a critical component of presenting, analyzing, and interpreting results. For instance, related to research questions two and three below, it is the graphics that will be most revealing in the differences, if any, and not the numerical analyses (though numerical analyses are provided).

Given that charter schools are regularly being offered as solutions for needed educational reform nationally, it is imperative that they be evaluated from a national perspective. This study proposes to compare the academic performance in two domains of charter and traditional public schools using the National Assessment of Educational Progress (NAEP) using multiple propensity score methods. More specifically, this study proposes to address the questions:

1. Given appropriate adjustments based on available student data, is there a discernible difference between charter and traditional public schools with regard to math and reading scores evaluated at grades 4 and 8?

2. If so, what is the nature and magnitude of this difference for the two outcomes, reading and mathematics?

3. And finally, what is the impact, if any, of different charter school laws?

CHAPTER 2: REVIEW OF THE LITERATURE

Though Ray Budde is often credited with the current charter school movement (Kolderie, 2005), the term *school choice* can be traced back to Adam Smith's *Wealth of Nations*, Thomas Paine's *Rights of Man*, and John Stuart Mill's *On Liberty* (Herbst, 2006). Prior to the Revolutionary War, given the religious diversity of colonial America, issues of education were left to local communities. However, after the war Revolutionary leaders argued that local schools were no longer sufficient for educating students for the emerging state and federal governments. It was Thomas Jefferson who, in 1779, introduced the first bill in Virginia that would establish a public school system. It was this, along with numerous other American intellects during the 1780s and 1790s, that established public schools throughout the young nation thereby relegating school choice to a choice between the public school and, predominately religious, private schools.

In the wake of the landmark report *A Nation at Risk* (The National Commission on Excellence in Education, 1983), Budde (1988) authored a pivotal document that started the charter school movement in the United States. In this document, Budde argues that system-wide changes to the way schools are structured are required including: more rigorous curriculum and graduation standards; extended school days and year; more homework; teacher accountability for student results; termination of "incompetent" teachers; and higher pay for teachers. To achieve these goals, he proposed a fundamental change to the "internal organization of the school district... making substantial changes in the roles of teachers, principals, the superintedent, the school board, parents, and others in the community" (p. 16). More specifically, a framework for charter schools was proposed that includes five stages over a three year period (see Figure 2). The five stages include: (1) generating ideas; (2) planning the charter; (3) preparing for teaching; (4) teaching under the educational charter; and (5) program monitoring and evaluation. For the first iteration of the cycle, stages one, two, and three occur

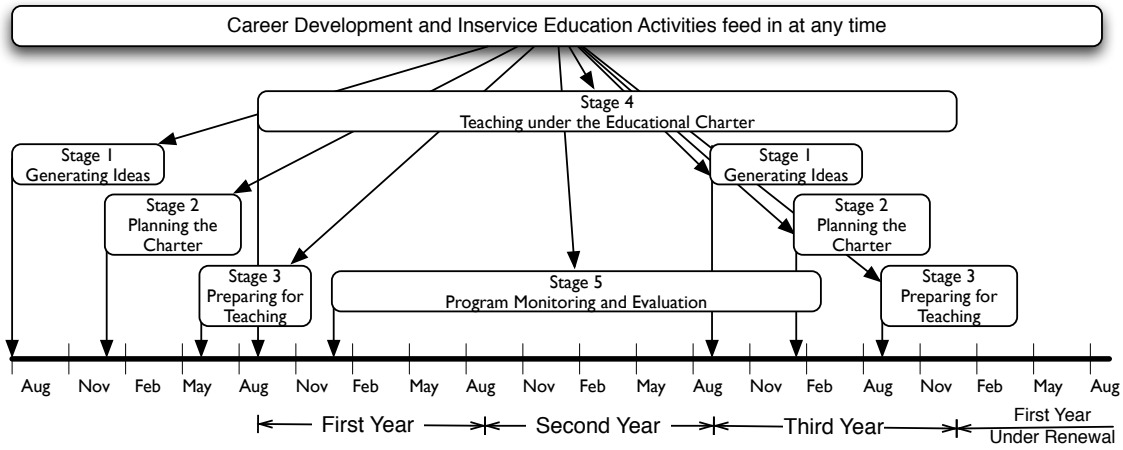


Figure 2: Stages of a Charter School Life Cycle (adapted from Budde, 1988)

prior to the opening of the school with stage one ideally beginning a full school year before. There are several features of this framework that deviate from traditional public school models, but most notably is the repetition of what may appear to be preparational stages. That is, the charter school must re-plan their school structure periodically (every three to five years according to Budde's framework) in a manner consistent to the initial charter school creation, thereby forcing a re-evaluation of the school bureaucracy.

Following the suggestions of Budde, Minnesota passed the first charter school law in 1991 with California being the second following in 1992. As of spring 2009, 40 states and the District of Columbia have charter school laws which comprise 1,407,421 students in 4,578 schools (Center for Education Reform, 2010). According to National Alliance of Public Charter Schools (2009), there are currently over 200 studies that examine charter school achievement.

Empirical Evidence for Charter School Effectiveness

Given that program evaluation and accountability are fundamental components of the philosophical foundations of charter schools, there are remarkably few *high quality* empirical studies that address, at a national level, the academic effectiveness of charter schools (c.f. National Alliance of Public Charter Schools, 2009; Betts &

Tang, 2008). That is not to say that there are not any studies that examine charter school achievement. The National Alliance of Public Charter Schools (2009) provide a review of 140 studies selected on several criteria. Their review reveals significant gaps in the research with regard to states evaluated, research quality that addresses achievement, as well as timeliness of results. This is further exemplified by a meta-analysis conducted by Betts and Tang (2008) that includes just 13 studies that represent nine states. In this section we will provide an overview of the current literature available vis-à-vis the published meta-analysis and literature reviews. We then focus on two recent studies that together provide the context for the study proposed here. More specifically, a hierarchical linear modeling analysis of the 2005 NAEP study (Braun et al., 2006a) and a matching study comparing charter and public schools in 2009 with 16 states (Center for Research on Education Outcomes, 2009) and again in 2013 with 27 states (Center for Research on Education Outcomes, 2013).

Overview of Current Studies

Research has shown that parents of students in charter schools are generally more satisfied with the charter school than the public school and will also tend to be more involved in their child's education (Teske & Schneider, 2001; Vanourek, Manno, Finn, & Bierlein, 1998). However, their satisfaction may simply be a rationalization (Hubbard & Kulkarni, 2009). Moreover, Fuller et al. (1996, as cited in Hubbard & Kulkarni, 2009) suggest that parents that choose charter schools "believe that the charter must therefore be superior to a conventional public school" (p. 177). This is collaborated by a study conducted by Cullen, Jacob, and Levitt (2005) that examines school choice in Chicago Public Schools whereby more than half of students elect to attend another Chicago Public School (e.g. career academy, high-achieving school) rather than their locally assigned school. Though students who opt out of their local school are more likely to graduate, Cullen et al. (2005) argue that "those who opt out are superior along unobservable dimensions such as

their motivation level and parental involvement" (p. 755).

The National Alliance of Public Charter Schools (2009) provides perhaps the most comprehensive review of available research on charter school performance. The current report, *Charter School Achievement: What We Know* is now in its fifth edition having been updated periodically to account for recent studies. In addition to covering published research reports, the review includes unpublished reports including conference presentations, dissertations, policy group and think tank reports, and state evaluations. Of the 210 studies identified, 140 are included in their review given that the study compares charter schools with traditional public schools, the study uses "serious research methods" (p. 2), and "examines a significant segment of the charter sector." The studies are then further categorized into one of three categories: (1) panel studies that are longitudinal and examine student growth over time; (2) cohort change studies that are longitudinal but use some other method than tracking individual students; and (3) snapshot studies that examine school performance at a single point in time (also known as observational studies).

Table 1 summarizes the findings of the 140 studies included first, by breaking out the year(s) the study's data is based upon, and second by the results reported. It should be noted that many of the pre-2001 studies were concentrated in a few states (Arizona, California, Florida, North Carolina, and Texas). This is expected given that these states are among the earliest to adopt charter school laws (see appendix A) as well as the substantial increase in the number of charter schools since 2000 (see figure 1). The National Alliance of Public Charter Schools conclude that:

> [I]t becomes dramatically clear that studies examining public charter schools in more recent academic years show that charter schools produce more instances of larger achievement gains in both math and reading when compared to traditional public schools. (p. 3)

However, this interpretation downplays the fact that approximately 30% of charter

Table 1: Summary of Studies on Charter School Achievement

	Pre 2001				Post 2001			
	Larger Gains	Similar Gains	Mixed Gains	Smaller Gains	Larger Gains	Similar Gains	Mixed Gains	Smaller Gains
Math	4	4	4	20	17	17	1	14
	(13%)	(13%)	(13%)	(63%)	(35%)	(35%)	(2%)	(29%)
Reading	7	10	3	14	18	12	1	14
	(21%)	(29%)	(9%)	(41%)	(40%)	(27%)	(2%)	(31%)

Source: National Alliance of Public Charter Schools, 2009

schools performed worse than their traditional public school counterpart. These results are consistent with a recent by the Center for Research on Education Outcomes (2009) that reported that 37% of charter schools performed worse than their public school counterpart in 2009 and 31% in 2013 (this study is discussed in more detail below).

Betts and Tang (2008) employ more stringent selection criteria for including studies in their meta-analysis. More specifically, only studies that used experimental student-level growth-based methods were included, resulting in a total of 14 studies published between 2001 and 2007 utilizing data ranging from 1998 through 2005. Similarly to National Alliance of Public Charter Schools (2009), studies included a limited number of locations including Arizona, California (three of which from San Diego specifically), Chicago, Delaware, Florida, Idaho, North Carolina, and Texas with one additional anonymous location. Overall, their analysis of the available studies provide very mixed results. However, some patterns to charter and public school differences emerge, specifically that charter schools generally outperform traditional public schools in elementary school reading and middle school math, though effect sizes for the latter are small. However, for high school reading and math charter schools are generally underperforming traditional public schools, but it should be noted that studies examining these grade levels is relatively small (see also, National Alliance of Public Charter Schools, 2009).

Two NAEP Studies Using HLM

An increasingly used statistical method that allows for the analysis of studies where observations are not independent is hierarchical linear modeling (HLM), or multi-level analysis. In the context of the charter school question, comparing students in charter schools to public school counterparts with, say ordinary least squares or ANOVA, is inappropriate since these statistical models do not account for the school effects. HLM provides a model for which school effects can be partitioned from student effects thereby providing adjustments for the lack of independence of observations (see e.g., Bryk & Raudenbush, 1992; Gelman & Hill, 2006).

Braun, Jenkins, and Grigg have published two research reports utilizing NAEP and HLM that look at how public school students compare to private school students (2006b) and charter schools students (2006a). Note that the former study used the 2005 administering of NAEP whereas the latter used the 2003 administering of NAEP. A key advantage of using NAEP is that many student (see appendix C) and school level variables are available. Moreover, as of 2003 charter schools have been oversampled to ensure sufficient sample sizes for appropriate comparisons to be made.

Comparing Private and Public Schools.

For the private school study (Braun et al., 2006b), results found that students in private schools scored significantly higher than public school students in both mathematics and reading at grades 4 and 8. Differences ranged from 8 points for grade 4 mathematics to 18 points for grade 8 reading. Adjusting for student characteristics with HLM resulted in reductions in all four comparisons of approximately 11 to 14 points. After adjustment, private school students still scored significantly higher than public school students in grade 8 reading, but public schools scored significantly better in grade 4 mathematics. There was no significant difference for grade 4 reading and grade 8 mathematics.

Comparing Charter and Public Schools.

For the charter school study (Braun et al., 2006a) analysis was conducted in three phases for both reading and mathematics. In phase one, all charter schools were compared to all public schools. Results found that, when student characteristics were adjusted for, charter schools performed on average 4.2 points lower than publics schools in reading (corresponding effect size is 0.11 standard deviations) and 4.7 points lower in mathematics (corresponding effect size is 0.17 standard deviations).

Phase two separated charter schools into two groups: charter schools that are associated with a public school district (PSD) and those that are not. Separate analysis were performed for each charter school type with public schools. For reading, there was no significant difference between charter schools affiliated with a PSD and public schools. However, for schools not affiliated with a PSD, charter school students scored significantly lower than public school students with an adjusted difference of 0.17 standard deviations. Similarly for mathematics, there was no difference between charter schools affiliated with a PSD and public schools but charter schools not affiliated with PSD scored significantly lower with an adjusted difference of 0.23 standard deviations.

Lastly, phase three compared only charter and publics schools located in a central city and serving a high-minority population. For reading, there was no significant difference between charter and public schools for any model. For mathematics however, charter schools not affiliated with a PSD scored significantly lower than public school students with an adjusted difference of 0.17 standard differences. There was no difference for schools affiliated with a PSD.

The CREDO Study

The Center for Research on Education Outcomes (2009, 2013) conducted a study of more than 1.7 million records from 2400 charter within 16 and 27 states in

2009 and 2013, respectively. The methodology involves creating a Virtual Control Record (VCR) for each charter school student (see also, Abadie, Diamond, & Hainmueller, 2007; Northwest Evaluation Association, 2009) which is used to find matching student from an eligible traditional public school. Students within a traditional public school become available in a pool of potential matches when at least one student is identified as transferring to a charter school. Once the "feeder schools" are identified, all students from feeder schools are pooled and serve as the source to select matches to the charter school students. Students are then matched on the following factors: grade-level, gender[2], race/ethnicity, free or reduced price lunch status, English language learner status, special education status, and prior test score on state achievement tests. This procedure, which is similar to propensity matching, resulted in 83.7% and 84.4% of charter school students being matched to a public school student for reading and math, respectively.

Once matches were determined, ordinary least squares regression was utilized to analyze both math and reading scores, separately, across the charter school students and matched public school students. Moreover, controls for student characteristics used above, excluding gender, along with state indicators and scores affected by Hurricane Katrina, were added to the basic model. Overall results show that charter school students performed, on average, 0.01 and 0.03 standard deviations below public school students for reading and math, respectively. Both results are significant at $p \leq 0.01$.

Though the magnitude of the overall effects may not necessarily suggest charter schools are performing substantially lower than their public school counterparts, further analysis by Center for Research on Education Outcomes (2009) reveal more nuanced understanding of the differences. More specifically, the effectiveness of charter schools varied considerably by state. Five states (Arkansas, Colorado, Illinois, Louisiana, & Missouri) were found to have higher learning gains for charter schools. Six states (Arizona, Florida, Minnesota, New Mexico, Ohio, & Texas) were

[2]Gender was not available in Florida

found to have lower learning gains for charter schools. The remaining four states (California, District of Columbia, Georgia, & North Carolina) had either mixed results or no difference in academic gains.

Lastly, the Center for Research on Education Outcomes (2009) found variation of charter school effectiveness across school characteristics. That is, schools that focused on elementary or middle grades separately, tended to perform as well or better than their public school counterparts. However, for charter schools that focused on high grades or multi-level grades performed anywhere from .02 to .08 standard deviations below public schools. Moreover, school level comparisons find that only 17% of charter schools perform better than public schools while 46% perform no differently and 37% perform significantly worse.

The results from the 2013 study (Center for Research on Education Outcomes, 2013) show a small increase in the difference between charter and traditional public school students. In 2009, charter school students had a loss of 7 days which increased to a gain of 8 days in 2013. For math, charter school students has a loss of 22 days in 2009 and are on par with traditional public school students in 2013. It should be noted that the Center for Research on Education Outcomes prefers to present their results in a metric of days. This is problematic since it is difficult to compare to other studies. However, these differences are approximately as effect sizes between 0.01 and 0.03 (Loveless, 2013). These results, as will be shown in chapter four, are consistent with the results of this study. Moreover, this also demonstrates an issue with null hypothesis testing and p-values, especially with large n's. This will be discussed in detail in chapter five.

Propensity Score Analysis

Randomized experiments are the gold standard for estimating causal effects of a treatment. However, as is frequently the case in educational contexts, randomization for the current project is neither ethical nor feasible. Therefore, propensity score methods (Rosenbaum & Rubin, 1983) using matching (Stuart &

Rubin, 2008; Stuart, 2010) and stratification methods (Raudenbush, Hong, & Rowan, 2003) will be used to make available quasi-experimental estimates of causal effects (see e.g. Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007; Stuart & Rubin, 2008). Recent research comparing the use of propensity score methods with randomized experiments have shown that causal estimates from observational studies using propensity score methods are generally consistent with those from randomized experiments (Cook, Shadish, & Wong, 2008; Shadish, Clark, & Steiner, 2008). The use of propensity score methods in published psychology and education research has been growing over the last decade (Thoemmes & Kim, 2011).

The selection of covariates is particularly challenging in propensity score analysis. As such, multiple methods for the estimation of propensity scores will be used (Rosenbaum, 2012). The goal in the estimation of propensity scores is to reduce selection bias, therefore simple significance testing is not appropriate (Rosenbaum, 2002) since potentially non-significant covariates may be proxies for important non-observed covariates. Although this study has been designed to measure the same covariates as would be used in a randomized block design, multiple methods utilizing varying number of covariates will be used. Moreover, we wish to provide overall effect sizes but also measure the effects of clustering of state.

Although propensity score analysis has been shown to provide estimates consistent with randomized experiments (Shadish et al., 2008; Dehejia & Wahba, 1999; Heckman, Ichimura, Smith, & Todd, 1997), its use has not been immune to criticism (c.f. Shadish, 2013). Pearl (2009) has raised concerns about a potential increase in bias due to the inclusion of certain covariates in the estimation of propensity scores in response to Rosenbaum's (2002) suggestion that in general the inclusion of all observable covariates is preferable to excluding them. However, Pearl's concerns can be mitigated in at least three ways. First, careful checking of balance across all observable covariates is done even if the covariate is not used in the modeling of the propensity scores. Second, sensitivity analysis (Rosenbaum,

2002) can be conducted after matching which considers the question of whether the estimate would differ in the presence of additional unobserved covariates. That is, this method tests the robustness of the propensity score estimation for hidden bias. Currently sensitivity analysis is only well defined for one-to-one matching and therefore is not used for this project. And third, utilize multiple methods for estimating propensity scores (Rosenbaum, 2012). Specifically, in addition to matching based upon propensity scores estimated from logistic regression, stratification methods using both quintiles on the logistic regression estimated propensity scores and classification trees provide parametric and non-parametric estimates, respectively.

The use of propensity score methods is still preferable over traditional regression models in light of these criticisms. Specifically, propensity score analysis separates the covariates related to selection bias and the comparison of the outcome of interest. This clean separation of the research design also provides a more clear interpretation of results similar to randomized experiments. As will be discussed more fully in chapter three, special must be placed on achieving balance. In the context of propensity score analysis, balance refers to reducing bias or differences between observed covariates for the units that will be compared. For example, if after matching we find that each matched pair has the same ethnicity we would conclude that perfect balance has been achieved for that covariate. In chapter three I will outline a number of approaches for checking balance for both matching and stratification methods. Although there is the possibility of a lack of balance in an unobserved covariate (i.e. "hidden bias"), this would similarly affect regression methods and is admittedly an important limitation of any non-randomized method for estimating causal effects. However, NAEP is designed to include most, if not all, the important covariates one would expect to be related to charter school attendance.

CHAPTER 3: METHOD

This chapter will outline the methods that will be utilized to describe and analyze the data in order to address the research questions central to this study. Given the strong political interests in the question of charter school effectiveness and the implications for educational policy both at the state and national level, obtaining good empirical evidence, preferably with strong causal inferences, is most desirable. The *gold standard* of inferential research is the randomized experiment. A research design that addresses the charter school question proposed here would require that students be randomly assigned, possibly with blocking on key covariates, to either a charter or public school. The theoretical justification for such a scheme is that any systematic differences between the two groups would be balanced through the randomization processes. However, in practice, especially in education, such randomization is neither feasible nor ethical. The result of the lack of randomization is a phenomenon called selection bias. That is, any comparisons of the two groups will be biased given the fact that the units of study, students in this study, self-selected to be in their respected group. Propensity score analysis (Rosenbaum & Rubin, 1983) is a statistical approach whereby the observed differences between the two groups are balanced by the careful analysis of covariate information. This procedure lends itself well to secondary analysis of observational data.

Overview of NAEP

The source of the data that will be utilized in this study is provided by the National Center for Educational Statistics (NCES) which is within the U.S. Department of Education's Institute of Education Sciences (IES). The National Assessment of Educational Progress (NAEP) was started in 1971 and has provided national measures of student achievement in many subjects including mathematics, reading, science, writing, history, civics, and the arts. In 2003 NAEP began

assessing charter schools as well as private and public schools. This study will utilize the 2007 administering of the NAEP assessments in mathematics and reading within grades four and eight. The 2007 assessment included over 6,000 public schools and over 200 charter schools comprising of over 145,000 and 3,000 students, respectively. Given this relatively large, nationally representative sample, analysis of NAEP assessments utilizing propensity score analysis may prove to provide valuable insights into the academic differences between charter and public schools.

More than simply providing large samples, another key advantage of NAEP is the fact that it is not designed to assess individual students or schools, but instead is designed to inform subject-matter achievement, instructional experiences, and school environments (Braun, Jenkins, & Grigg, 2006). To achieve this goal, NAEP utilizes a complex item-sampling design such that individual students are presented a subset of the total items, thereby reducing the burden on participants. Though not appropriate for assessing individual student achievement, in aggregate the NAEP measures provide a robust and accurate estimate of student achievement.

In addition to subject area measures, NAEP includes student, teacher, and school questionnaires that provide contextual information about the students' environment. Given that PSA relies on adjusting for selection bias by adjusting for known covariates, it are the answers to these questionnaires that will serve as the basis for determining a students propensity score, or likelihood of being in the treatment (i.e. charter school in the context of this study). In addition to typical demographic items such as gender and race, students are asked about computers, books, magazines, and encyclopedias in the home; parents education level; and the level of interaction with academics within the home (see appendix C for complete list of items).

The responsibility for developing the assessment objectives and test specifications lies with the National Assessment Governing Board which was created by Congress in 1988. Traditionally they are the states that have provided the legal

Table 2: Distribution of Math Items by Grade and Content Area

Content Area	Grade 4	Grade 8
Number Properties and Operations	40%	20%
Measurement	20%	15%
Geometry	15%	20%
Data Analysis and Probability	10%	15%
Algebra	15%	30%

guidance for school governance including accountability measures. Given the varied standards across states, it is this governing board that is to determine nationally what are appropriate achievement goals for each age and grade. The following two sections will provide the framework for mathematics and reading assessments.

Mathematics

Since 1990, the Council of Chief State School Officers (CCSSO) has been contracted to design a framework for the mathematics assessment (National Assessment Governing Board, 2006a). The framework was most recently updated in 2000 to take into account state standards, the National Council of Teachers of Mathematics (NCTM) standards, the Trends International Mathematics and Science Study (TIMSS), the Achieve Project, and a 2001 report issued by the National research Council of the National Academy of Sciences. The result of their work was six recommendations for the mathematics assessment regarding content areas, mathematical complexity of items, distribution of items, item formats, manipulatives, and calculators. For the purposes of the study proposed, a composite score will be utilized that is comprised of five content areas, number properties and operations; measurement; geometry; data analysis and probability; and algebra. Table 2 provides details regarding the distribution of items comprising the composite score for the grade four and eight assessments.

Reading

The NAEP Reading Framework (2006) provides guidelines and a theoretical basis for reading assessment. This framework is designed with the input of individuals and organizations involved in reading education including researchers, policymakers, teachers, and business representatives. However, a particular emphasis is placed on the work of the National Institute for Child Health and Human Development (NICHID). More specifically, the NICHID summarizes how the research describes a reader as:

> In the cognitive research, reading is purposeful and active. According to this view, a reader reads a text to understand what is read, to construct memory representations of what is understood, and to put this understanding to use. (p. 4, NICHD, 2000, as cited in National Assessment Governing Board, 2006b)

Moreover, reading is considered to be a complex process rather than a simple set of skills. As such, the NAEP reading assessment is designed such that comprehension is defined as:

> "[I]ntentional thinking during which meaning is constructed through interactions between text and reader" (Harris & Hodges, 1995). Thus, readers derive meaning from text when they engage in intentional, problem solving thinking processes. (p. 14, NICHD, 2000, as cited in National Assessment Governing Board, 2006b)

Given this framework, NAEP provides an excellent tool for evaluating overall reading achievement, but not to diagnosis specific individuals.

The NAEP reading assessment is designed to account for three reading contexts: reading for literacy experience, reading for information, and reading to perform a task. Within these contexts, four aspects of reading are considered: forming a general understanding, developing interpretation, making reader/text connections, and examining content and structure. The reading assessment is administered by supplying students with booklets that contain reading materials and comprehension questions. The questions consist of both multiple-choice and constructed-response question formats with at least half of the questions being of the constructed-response type.

Analysis

This study utilizes propensity score analysis for estimating causal effects of students attending charter schools. The propensity score is "the conditional probability of assignment to a particular treatment given a vector of observed covariates" (Rosenbaum & Rubin, 1983). The probability of being the in treatment is defined as:

$$\pi(X_i) \equiv Pr(T_i = 1|X_i)$$

Where X is a matrix of observed covariates. The balancing property under exogeneity states that,

$$T_i \perp\!\!\!\perp X_i \mid \pi(X_i)$$

In the case of randomized experiments, the strong ignobility assumption states,

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i|X_i$$

for all X_i. That is, treatment is independent of all covariates, observed or otherwise. However, we can restate the strong ignobility assumption with the propensity score as,

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i \mid \pi(X_i)$$

So that treatment placement is ignorable given the propensity score presuming sufficient balance[3] is achieved.

The average treatment effect (ATE) is defined as $E(r_1) - E(r_0)$ where $E(.)$ is the expected value in the population. Given a set of covariates, X, and outcomes Y, where 0 denotes traditional public school student and 1 denotes charter school

[3]Balance in the context of PSA refers to differences in observed covariates between treatment and control units is minimized.

student, ATE is defined as:

$$ATE = E(Y_1 - Y_0|X) = E(Y_1|X) - E(Y_0|X)$$

Or the difference between charter and traditional public school given the set observed covariates.

Rosenbaum (2012) suggests that hypotheses should be tested more than once in observational study. This study will estimate treatment effects using nine separate propensity score models within three larger classes. The third class of PSA, multilevel PSA, and as implemented in the `multilevelPSA` R package, was developed in part for this dissertation. For each of the four subject and grade combinations, the following methods will be used resulting in a total of 36 propensity score analyses being conducted.

1. Propensity score analysis using stratification. This method ignores state assignment as a clustering variable. Under this broader method three statistical methods for stratification will be used:

   (a) Full logistic regression. This method will estimate propensity scores using logistic regression with all available covariates, but will exclude interaction or product terms.

   (b) Logistic regression with step AIC. The step AIC in the MASS package (Venables & Ripley, 2002) will select the best logistic model based upon the Akaike Information Criterion (Akaike, 1974). In this case the best first order interaction terms will be added to the main effect terms in a.

   (c) Conditional inference trees, based on all covariates; missing data will also be accommodated with the tree-based methods.

2. Propensity score matching. This method implicitly accounts for clustering. That is, the method used will find matches between treated and control units

that first match exactly on state, ethnicity, and gender, then finds a best match based upon the propensity scores estimated using logistic regression. As suggested by Stuart (2010), multiple matched sets will be formed using:

(a) One-to-one (i.e. one charter school student is matched to no more than one traditional public school student).

(b) One-to-five (i.e. one charter school student is matched to as many as five traditional public school students).

(c) One-to-ten (i.e. one charter school student is matched to as many as ten traditional public school students).

A dependent sample analysis will be performed on the resulting matched pairs (Austin, 2011).

3. Multilevel propensity score analysis (see e.g. Bryer, 2011). This method will utilize the same stratification methods as described in method one above, namely:

(a) Full logistic regression.

(b) Logistic regression with step AIC.

(c) Conditional inference trees.

Graphical Representation

Given the large amount of data to be summarized, the use of graphics will be an integral component of representing the results. The `multilevelPSA`[4] package provides a number of graphing functions that extend the framework introduced by Helmreich and Pruzek (2009) for multilevel PSA. Figure 3 represents a multilevel PSA assessment plot with annotations. This graphic represents the results of

[4]The `multilevelPSA` package was written by the author and is available from `http://github.com/jbryer/multilevelPSA`.

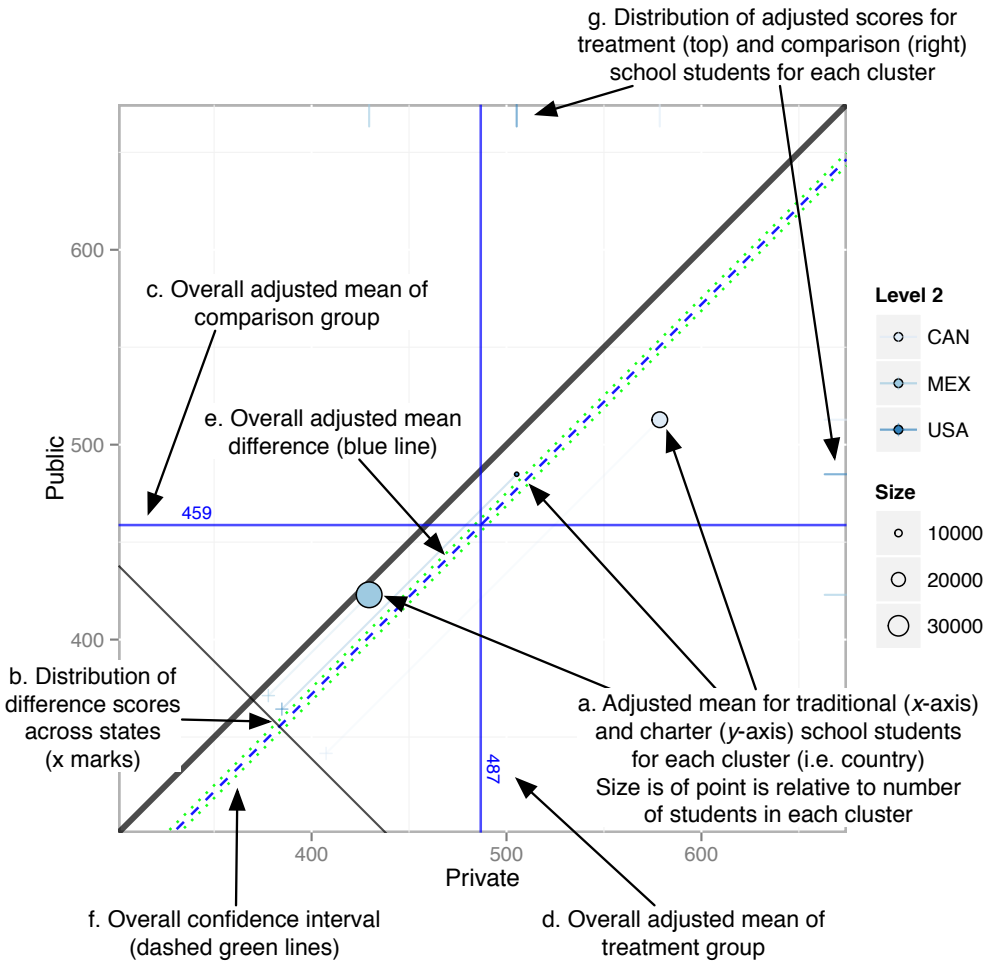


Figure 3: Annotated Multilevel PSA Assessment Plot

comparing private and public schools in North America using the Programme of International Student Assessment (PISA; Organisation for Economic Co-Operation and Development, 2009). The PISA data to create this graphic is included in the `multilevelPSA` package and a more detailed description of how to create this graphic are discussed at the end of this chapter. The following section will focus on the features of this graphic.

In Figure 3, the x-axis corresponds to math scores for private schools and the y-axis corresponds to public school maths cores. Each colored circle (a) is a country

with its size corresponding the number of students sampled within each country. Each country is projected to the lower left, parallel to the unit line, such that a tick mark is placed on the line with slope -1 (b). These tick marks represent the distribution of differences between private and public schools across countries. Differences are aggregated (and weighted by size) across states. For math, the overall adjusted mean for private schools is 487 and the overall adjusted mean for public schools is 459 and represented by the horizontal (c) and vertical (d) blue lines, respectively. The dashed blue line parallel to the unit line (e) corresponds to the overall adjusted mean difference and likewise, the dashed green lines (f) correspond to the confidence interval. Lastly, rug plots along the right and top edges of the graphic (g) correspond to the distribution of each country's overall mean private and public school math scores, respectively.

Figure 3 represents a large amount of data and provides much greater insight into the data and results. The figure provides overall results that would be present in a traditional table, for instance the fact that the green dashed lines do not span the unit line (i.e. y = x) indicates that there is a statistically significant difference between the two groups. However additional information is difficult to convey in tabular format. For example, the rug plots indicate that the spread in the performance of both private and public schools across countries is large. We can also observe that Canada, which has the largest PISA scores for both groups, also has the largest difference (in favor of private schools) as represented by the larger distance from the unit line.

The `multilevelPSA` R Package

All of the analysis for this study were conducted using R (R Development Core Team, 2008)., the use of which R provides a number a number of important advantages. First, all the analysis is reproducible. That is, researchers can

download all the R scripts [5] and those with access to the restricted NAEP data [6] can run all the analyses. Another important advantage of using R is that it is an extensible vis--vis R packages. Packages are collections of functions, data, and documentation designed for a specific purpose. Since the multilevel PSA methods described in this dissertation have never been conducted or implemented elsewhere, the `multilevelPSA` package was developed for R (R Development Core Team, 2008). As of this writing, version 1.2 is available on The Comprehensive R Archive Network (CRAN)[7]. In this section I will outline the core functionality of the `multlilevelPSA` package. Appendix J provides a complete list of the available functions with brief descriptions of their purpose. By convention, R commands are type faced in a fixed-width font and begin with a greater than (`>`) symbol.

To begin, the `install.packages` and `require` functions will install and load the package, respectively.

```
> install.packages('multilevelPSA', repos='http://cran.r-project.org')
> require('multilevelPSA')
```

The `multilevelPSA` package includes North American data from the Programme of International Student Assessment (PISA; Organisation for Economic Co-Operation and Development, 2009). This data is made freely available for research and is utilized here so that the R code is reproducible[8]. This example compares the performance of private and public schools clustered by country.

```
> data(pisana)
> data(pisa.psa.cols)
```

The `mlpsa.ctree` function performs phase I of the propensity score analysis using classification trees, specifically using the `ctree` function in the `party` package.

[5] Available on Github at `https://github.com/jbryer/Dissertation`.

[6] Typically data is included for the analysis to be fully reproducible. However, given the sensitive nature of the data the National Center for Education Statistics (NCES) requires a restricted license for access to the data.

[7] The CRAN package page is available at: `http://cran.r-project.org/web/packages/multilevelPSA/index.html`. The project source code is hosted on Github at: `https://github.com/jbryer/multilevelPSA`.

[8] NAEP requires a restricted use license and therefore the data is only available to qualified researchers. The R scripts for all analysis however, are available on Github at `http://github.com/jbryer/Dissertation`.

The `getStrata` function will return a data frame with a number of rows equivalent to the original data frame indicating the strata for each student.

```
> mlpsa <- mlpsa.ctree(pisana[,c('CNT','PUBPRIV',pisa.psa.cols)],
                       formula=PUBPRIV ~ ., level2='CNT')
> mlpsa.df <- getStrata(mlpsa, pisana, level2='CNT')
```

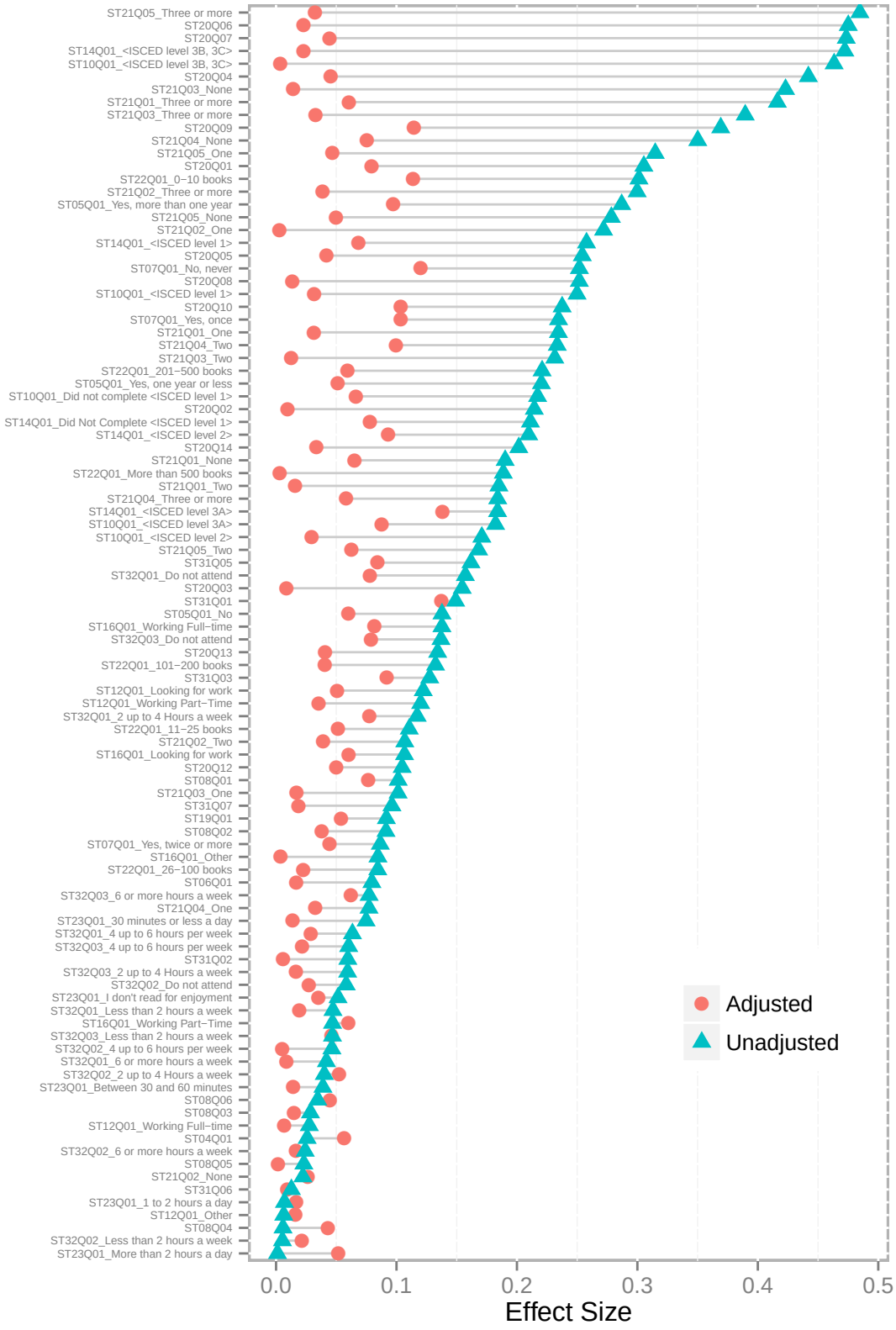
Similarly, the `mlpsa.logistic` estimates propensity scores using logistic regression. The `getPropensityScores` function returns a data frame with a number of rows equivalent to the original data frame

```
> mlpsa.lr <- mlpsa.logistic(pisana[,c('CNT','PUBPRIV',pisa.psa.cols)],
                             formula=PUBPRIV ~ ., level2='CNT')
> mlpsa.lr.df <- getPropensityScores(mlpsa.lr, nStrata=5)
> head(mlpsa.lr.df)
  level2    ps strata
1    CAN 0.917      2
2    CAN 0.941      3
3    CAN 0.969      4
4    CAN 0.930      2
5    CAN 0.836      1
6    CAN 0.973      4
```

The `covariate.balance` function will calculate balance statistics for each covariate by estimating the effect of each covariate before and after adjustment. The results can be converted to a data frame to view numeric results or the `plot` function will provide a balance plot.

```
> cv.bal <- covariate.balance(covariates=student[,pisa.psa.cols],
                              treatment=student$PUBPRIV,
                              level2=student$CNT,
                              strata=mlpsa.df$strata)
```

```
> head(as.data.frame(cv.bal))
  covariate es.adj es.adj.wtd es.unadj
1   ST04Q01 0.0565  -0.000396   0.0258
2   ST06Q01 0.0167  -0.000292   0.0796
3   ST08Q01 0.0766   0.000515   0.1014
4   ST08Q02 0.0379   0.000500   0.0913
5   ST08Q03 0.0150  -0.000850   0.0286
6   ST08Q04 0.0431  -0.000278   0.0058
> plot(cv.bal)
```


ST21Q05_Three or more
ST20Q06
ST20Q07
ST14Q01_<ISCED level 3B, 3C>
ST10Q01_<ISCED level 3B, 3C>
ST20Q04
ST21Q03_None
ST21Q01_Three or more
ST21Q03_Three or more
ST20Q09
ST21Q04_None
ST21Q05_One
ST20Q01
ST22Q01_0−10 books
ST21Q02_Three or more
ST05Q01_Yes, more than one year
ST21Q05_None
ST21Q02_One
ST14Q01_<ISCED level 1>
ST20Q05
ST07Q01_No, never
ST20Q08
ST10Q01_<ISCED level 1>
ST20Q10
ST07Q01_Yes, once
ST21Q01_One
ST21Q04_Two
ST21Q03_Two
ST22Q01_201−500 books
ST05Q01_Yes, one year or less
ST10Q01_Did not complete <ISCED level 1>
ST20Q02
ST14Q01_Did Not Complete <ISCED level 1>
ST14Q01_<ISCED level 2>
ST20Q14
ST21Q01_None
ST22Q01_More than 500 books
ST21Q01_Two
ST21Q04_Three or more
ST14Q01_<ISCED level 3A>
ST10Q01_<ISCED level 3A>
ST10Q01_<ISCED level 2>
ST21Q05_Two
ST31Q05
ST32Q01_Do not attend
ST20Q03
ST31Q01
ST05Q01_No
ST16Q01_Working Full−time
ST32Q03_Do not attend
ST20Q13
ST22Q01_101−200 books
ST31Q03
ST12Q01_Looking for work
ST12Q01_Working Part−Time
ST32Q01_2 up to 4 Hours a week
ST22Q01_11−25 books
ST21Q02_Two
ST16Q01_Looking for work
ST20Q12
ST08Q01
ST21Q03_One
ST31Q07
ST19Q01
ST08Q02
ST07Q01_Yes, twice or more
ST16Q01_Other
ST22Q01_26−100 books
ST06Q01
ST32Q03_6 or more hours a week
ST21Q04_One
ST23Q01_30 minutes or less a day
ST32Q01_4 up to 6 hours per week
ST32Q03_4 up to 6 hours per week
ST31Q02
ST32Q03_2 up to 4 Hours a week
ST32Q02_Do not attend
ST23Q01_I don't read for enjoyment
ST32Q01_Less than 2 hours a week
ST16Q01_Working Part−Time
ST32Q03_Less than 2 hours a week
ST32Q02_4 up to 6 hours per week
ST32Q01_6 or more hours a week
ST32Q02_2 up to 4 Hours a week
ST23Q01_Between 30 and 60 minutes
ST08Q06
ST08Q03
ST12Q01_Working Full−time
ST04Q01
ST32Q02_6 or more hours a week
ST08Q05
ST21Q02_None
ST31Q06
ST23Q01_1 to 2 hours a day
ST12Q01_Other
ST08Q04
ST32Q02_Less than 2 hours a week
ST23Q01_More than 2 hours a day
Adjusted
Unadjusted
0.0
0.1
0.2
0.3
0.4
0.5
Effect Size

The `mlpsa` function performs phase II of propensity score analysis and requires four parameters: the response variable, treatment indicator, strata, and clustering indicator. The `minN` parameter (which defaults to five) indicates what the minimum strata size is to be included in the analysis. For this example, 463, or less than once percent of students were removed because the strata (or leaf node for classification trees) did not contain at least five students from both the treatment and control groups.

```
> results.psa.math <- mlpsa(response=mlpsa.df$MathScore,
                            treatment=mlpsa.df$PUBPRIV,
                            strata=mlpsa.df$strata,
                            level2=mlpsa.df$CNT)
Removed 463 (0.696%) rows due to strata size being less than 5
```

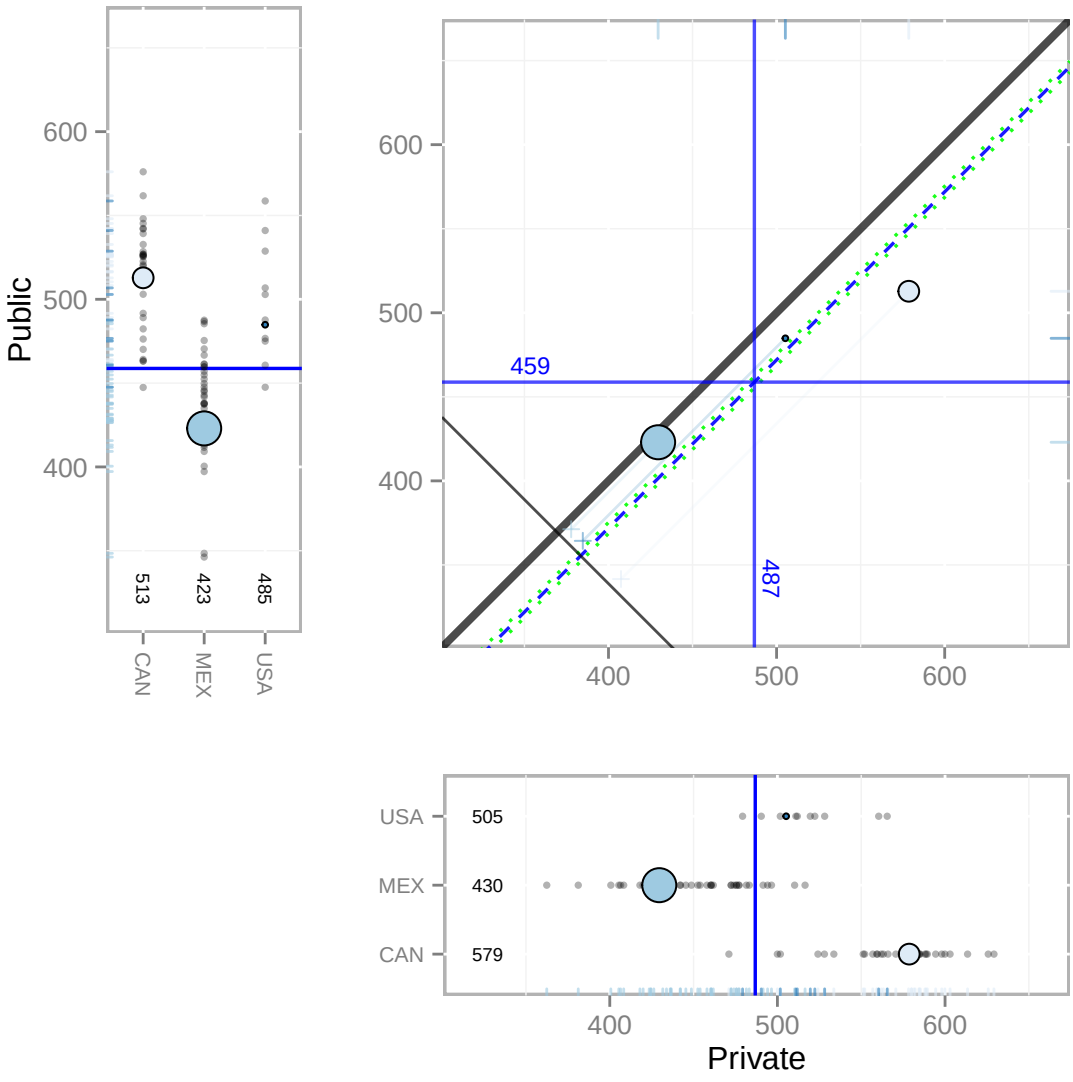
The `summary` function provides the overall treatment estimates as well as level one and two summaries.

```
> summary(results.psa.math)
Multilevel PSA Model of 85 strata for 3 levels.
Approx t: -10.8
Confidence Interval: -31.3, -24.75
   level2  strata Treat Treat.n Control Control.n ci.min ci.max
1     CAN Overall   579    1625     513     21093  -72.1 -59.57
2    <NA>       1   580      28     492      1128     NA     NA
3    <NA>       2   600       9     476      1326     NA     NA
... # Output truncated to save space
```

The `plot` function will create the multilevel assessment plot. Here it is depicted with side panels showing the distribution of math scores for all strata for public school students to the left and private school students below. These panels can be plotted separately using the `mlpsa.circ.plot` and `mlpsa.distribution.plot`

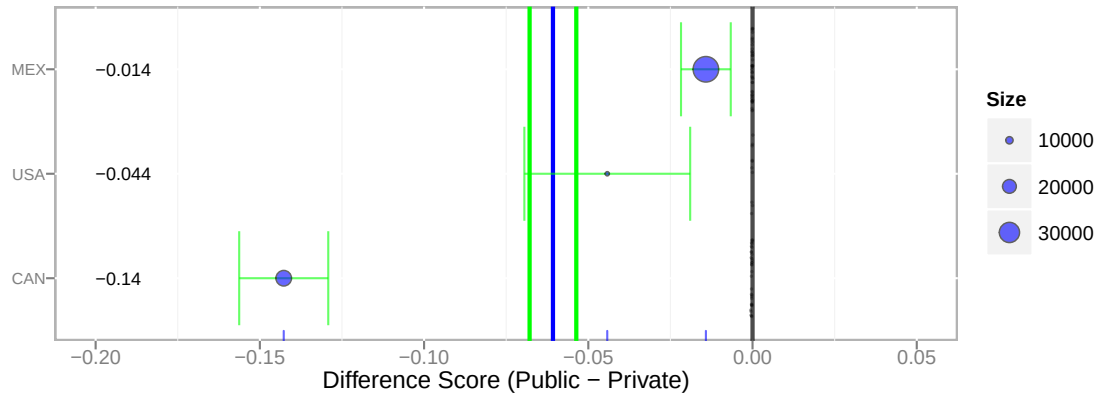
functions.

```
> plot(results.psa.math)
```



Lastly, the `mlpsa.difference.plot` function plots the overall differences. The `sd` parameter is optional, but if specified the x-axis can be interpreted as effect sizes.

```
> mlpsa.difference.plot(results.psa.math,
                        sd=mean(mlpsa.df$MathScore, na.rm=TRUE))
```



CHAPTER 4: RESULTS

This chapter will outline in detail the results of all the propensity score models described in chapter three. Since NAEP is organized such that each grade and subject combination is a separate dataset, this chapter will focus on the analysis of grade four math. The results for grade four reading, grade eight math, and grade eight reading are included in the appendices. The chapter begins with a discussion of data preparation, followed by details of the nine propensity score methods used, and concludes with a summary, including tables and figures, of the overall results across all grades and subjects.

All analysis was conducted using R (R Development Core Team, 2008)[9]. R provides a number of advantages over other applications including a framework for extending its core functionality through R packages. I have written and published two R packages primarily for conducting the analysis in this dissertation. The `naep` package provides functions to read and work with the National Assessment of Educational Progress (NAEP) data sets. Secondly, the `multilevelPSA` package provides functions to conduct multilevel propensity score analysis as described above. Both of these packages are available from the Comprehensive R Archive Network (CRAN).

Formatting note. Since the development of the R packages are in and of themselves a major component of this dissertation, I will make reference to some of the functions available. All references to R packages and functions will appear in a `fixed width font`.

Data Preparation

The propensity score methods I will use attempt to adjust for selection bias using the available observed covariates. However, the the geographic distribution of charter schools is not equal. That is, charter schools are more prevalent in certain

[9] All R scripts are available from Github at `https://github.com/jbryer/Dissertation`. Due to the licensing agreement with NCES, data is not included. However, researchers with access to the 2009 restricted use data should be able to replicate the analysis outlined in this chapter.

geographic regions of the country, often within urban areas. Since there are several orders of magnitude difference in the number of charter school to traditional public school students, selecting a subset of all the traditional public school students available in NAEP is possible. Specifically, by selecting traditional public school students who live in close proximity to a charter school the likelihood of them actually choosing a school increases. According to the National Household Travel Study, students travel an average of five miles to school. The Common Core of Data (National Center for Educational Statistics, 2009) provides the location of very public school in the United States. For each traditional public school the distance to the closest charter school was calculated using line-of-sight distance. Approximately 20% of traditional public schools, within states that have charter schools, were more than five miles from a charter school. Those schools, and subsequently students attending those schools and in any of the NAEP datasets, were eliminated from the study.

Table 3: Descriptive Statistics of Dependent Variables (Unadjusted) for All and Close (within 5 miles) Traditional Public Schools

	Charter	All Public Schools			Close Public Schools		
Subject	Mean	Mean	n	Diff	Mean	n	Diff
Grade 4 Math	231.5	238.3	159338	-6.9	237.3	85272	-6.0
Grade 4 Reading	212.9	218.8	168597	-6.0	217.1	92756	-4.2
Grade 8 Math	272.1	280.7	152048	-8.6	279.4	71528	-7.5
Grade 8 Reading	256.2	261.7	151304	-5.5	259.8	73810	-3.6

Appendix B provides descriptive statistics for all the covariates for the four datasets. Additionally, unadjusted differences in NAEP scores for each state containing a charter school are provided. Table 4 below provide the overall, unadjusted, differences in NAEP scorers for the four datasets.

Missing Data Imputation

Logistic regression, which is one of the two ways propensity scores will be estimated, require a complete dataset for estimation. Appendix D provides figures created using the `missing.plot` function in the `multilevelPSA` package

Table 4: Descriptive Statistics of Dependent Variables (Unadjusted)

	Charter		Public		Mean	Confidence	
Subject	Mean	SD	Mean	SD	Difference	Interval	
Grade 4 Math	231.2	28.3	237.3	28.5	-6.0	-7.0	-5.0
Grade 4 Reading	212.9	33.0	217.1	34.5	-4.2	-5.3	-3.1
Grade 8 Math	271.8	35.5	279.4	35.8	-7.5	-8.7	-6.4
Grade 8 Reading	256.2	32.9	259.8	32.6	-3.6	-4.7	-2.6

representing the extent of missingness for each covariate within each state. The first thing these figures reveal is that there is complete missingness in the majority of covariates for Alaska in grade four. As a result Alaska was removed from all datasets and will not be included in the study.

Secondly, the figures show that there are fewer than 5% of values are missing for the vast majority of covariates. In grade four math and reading, the three exceptions are newspapers in home, magazines in home, and encyclopedia in home. Grade eight math and reading also show a higher rate of missingness in these three covariates, but also in parent's education level. To examine whether data is missing at random, a logistic regression model was estimated predicting treatment from a shadow matrix (i.e. a matrix with the same dimension of the original dataset with 0s and 1s where 1s indicate the value is missing). The results of these models indicate that there is no relationship between charter school attendance and whether a student completed items regarding newspapers, magazines, and encyclopedias in the home. However, for grade eight math and reading, missingness of mother's and father's education level were statistically significant (p ¡ .05) predictors of treatment. It should also be noted that charter school students mother's education level was less likely to be missing whereas father's education level was less likely to be missing for traditional public schools. Although these two covariates are often important for understanding students educational backgrounds, the figures in Appendix I depicting the relative importance of each covariate for predicting charter school attendance using conditional inference trees which are estimated with missingness included, suggest that these covariates have relatively

low, or no, predictive value for charter school attendance. Therefore, missing values for these covariates will be imputed and used to estimate propensity scores for the logistic regression and matching methods.

Multiple imputation (Rubin, 1987, 1996) has become a popular approach for imputing missing values in datasets. For this study, missing data was imputed using by multivariate imputation by chained equations vis-à-vis the MICE package van Buuren and Groothuis-Oudshoorn (2011) in Rvan Buuren and Groothuis-Oudshoorn (n.d.). This package implements the fully conditional specification (FCS) method of imputation whereby separate multivariate imputation models are estimated for each variable containing missing values so that each model has its own set of conditional densities. Since the algorithm iterates through the data in small steps, providing diagnosing the imputed values as proceeding, the result is a robots estimate of imputed values. For this study, we will utilize both the original incomplete data for estimating propensity scores with classification trees and the complete imputed data for estimation propensity scores using logistic regression.

Propensity Score Analysis with Stratification

The first class of propensity score methods used is stratification. The general approach of stratification methods is to subdivide the available sample into smaller groups that have similar covariate profiles. Then a comparison using of mean differences between the treatment and control are made and an overall result is pooled from those individual comparisons. There are several ways to stratify the sample. For this study deciles based upon the propensity scores (i.e. fitted values of a logistic regression model) and leaves of a fitted classification tree. Moreover, given the importance of covariate selection and omission from propensity score models, two types of logistic regression models are used, namely a full model using all available covariates and an Akaike Information Criterion (AIC; Akaike, 1974) optimized model. The former is determined by a stepwise model selection algorithm where covariates are added and dropped and the model with that optimizes the AIC

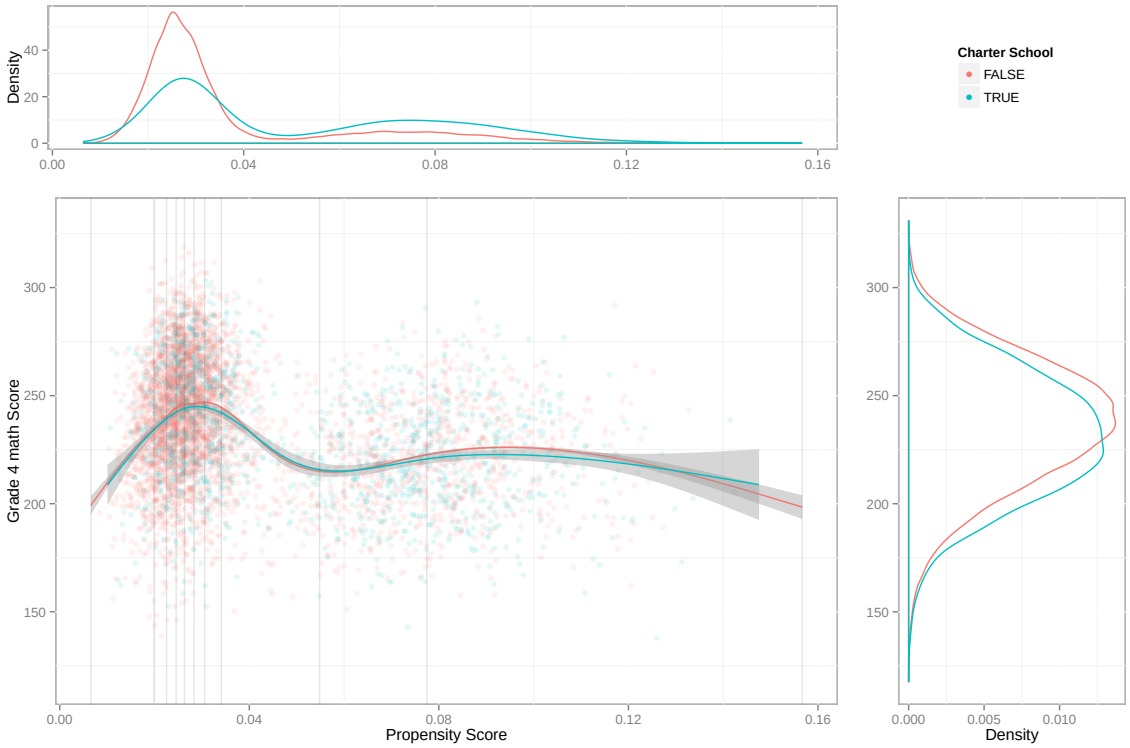


Figure 4: Loess Regression Assessment Plot: Grade 4 Math

is retained. Like all analysis in phase one, this is done without outcome variables.

However, before stratifying the logistic regression models we can examine the relationship between propensity scores and the outcome variable for the two groups. Furthermore, fitting a Loess regression line to that scatter plot provides an overall indication of the differences, if any. Figure 4 is a Loess Regression Assessment Plot created using the `loess.plot` function in the `multilevelPSA` package. The main panel is a scatterplot of each students propensity score on the x-axis and math score on the y-axis (for clarity a random 10% sample of data points are plotted). Two fitted Loess regression lines with approximate 95% confidence intervals are also plotted (Loess lines are based upon the full dataset and not the sample). The panel on the top provides a density distributions of the propensity scores and shows that there is generally good overlap between the two groups. Having adequate overlap is

critical since it indicates there are treatment and comparison units with similar propensity scores that will eventually will be compared on their outcome variables. The panel on the right is a density distribution of the unadjusted outcome and here shows that before propensity score adjustment traditional public school students performed slightly better than charter school students. However, given the strong overlap in the two Loess regression lines, this figure suggests there is no discernible difference in performance between traditional public school students and charter school students in grade four math. Corresponding plots for the other datasets as well as those for the AIC optimized models are provided in Appendix D.

The vertical lines in the main panel of Figure 4 represent the deciles, or strata. Figure 6 is a propensity score assessment plot where the x-axis is the outcome score for charter schools and the y-axis is the outcome sore for traditional public schools. Each circle corresponds to each strata and the size of the circle is proportional to the number of students within each strata. For the Logistic regression models, since deciles were used, each circle is of the same size. Figure 7 is the corresponding propensity score assessment plot for the classification tree model and therefore each strata is not of the same size. For points that lie on or near the unit line, $y = x$, indicate no significant difference in the outcome of the two scores. Lines are projected to a line perpendicular to the unit line and the tick placed. These tick marks correspond to the distribution of difference scores and the dashed blue line parallel to the unit line the overall mean difference. Furthermore, the green bar represents exactly the 95% confidence interval. Therefore, we can interpret the fact that the confidence interval does not span the unit line and is on the tradition public school side to indicate there is small statistically significant difference in favor of traditional public school students. Tables 5, 6, and 7 provide numeric results for each strata. Appendix F contains propensity score assessment plots and summary tables for grade four reading, grade eight math, and grade eight reading.

Covariate Balance

The goal of propensity score methods is to adjust for selection bias with the available observed covariates. In practice we test for the effectiveness of bias reduction by evaluating covariate balance. Perfect balance is achieved when there are no differences in covariate values for any matched pair or strata. However, perfect balance is almost never achieved. Figure 5 is a Covariate Effect Size balance plot. For each covariate on the y-axis the absolute standardized effect size before adjustment (in red) and after adjustment (in blue) are plotted. Effect sizes for individual strata are represented by letters. This figures shows that the propensity score adjustment greatly reduced the effects of each covariate. The remaining covariate balance plots are provided in Appendix E.

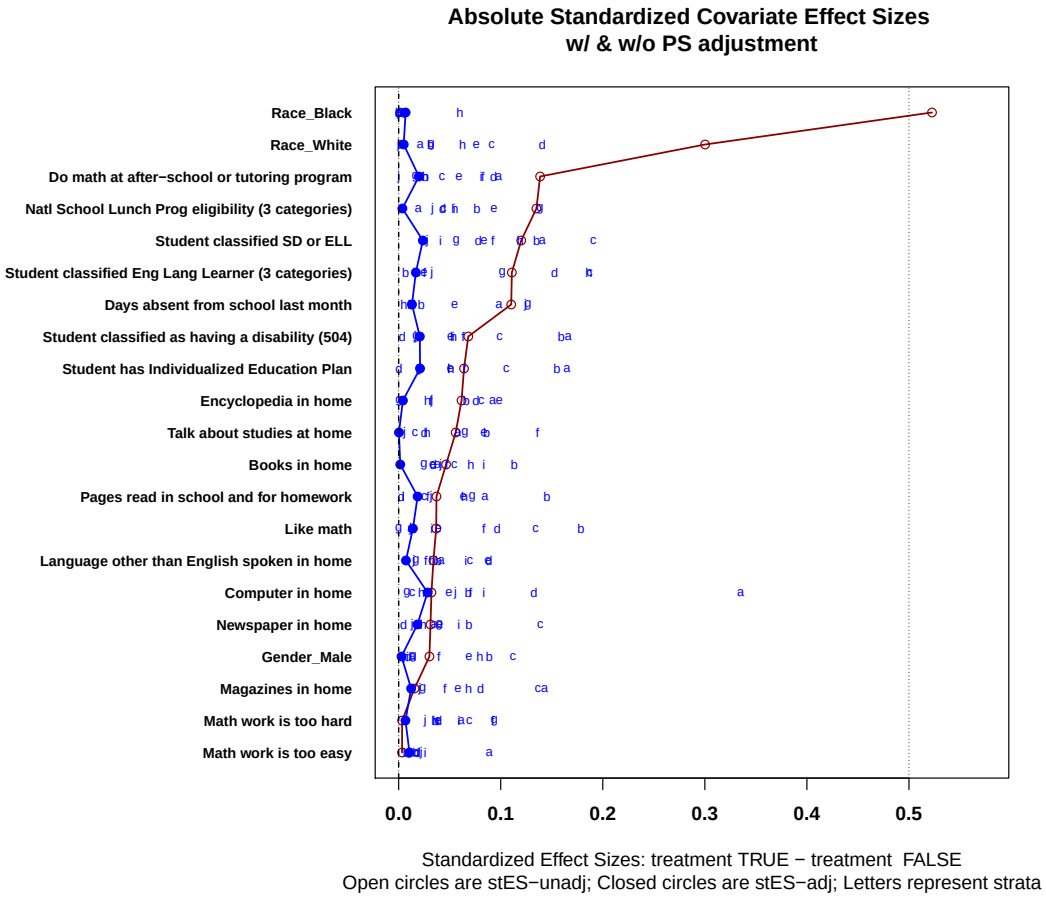


Figure 5: Covariate Balance Plot for Logistic Regression Stratification: Grade 4 Math

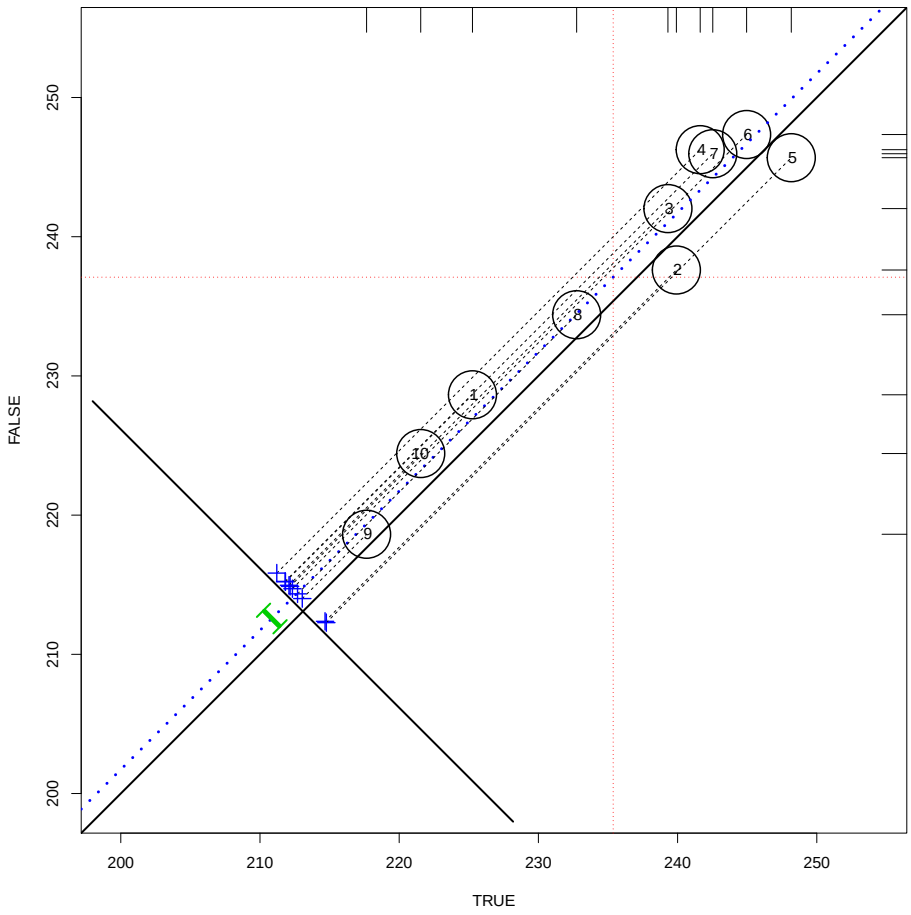


Figure 6: Propensity Score Assessment Plot for Logistic Regression Stratification: Grade 4 Math

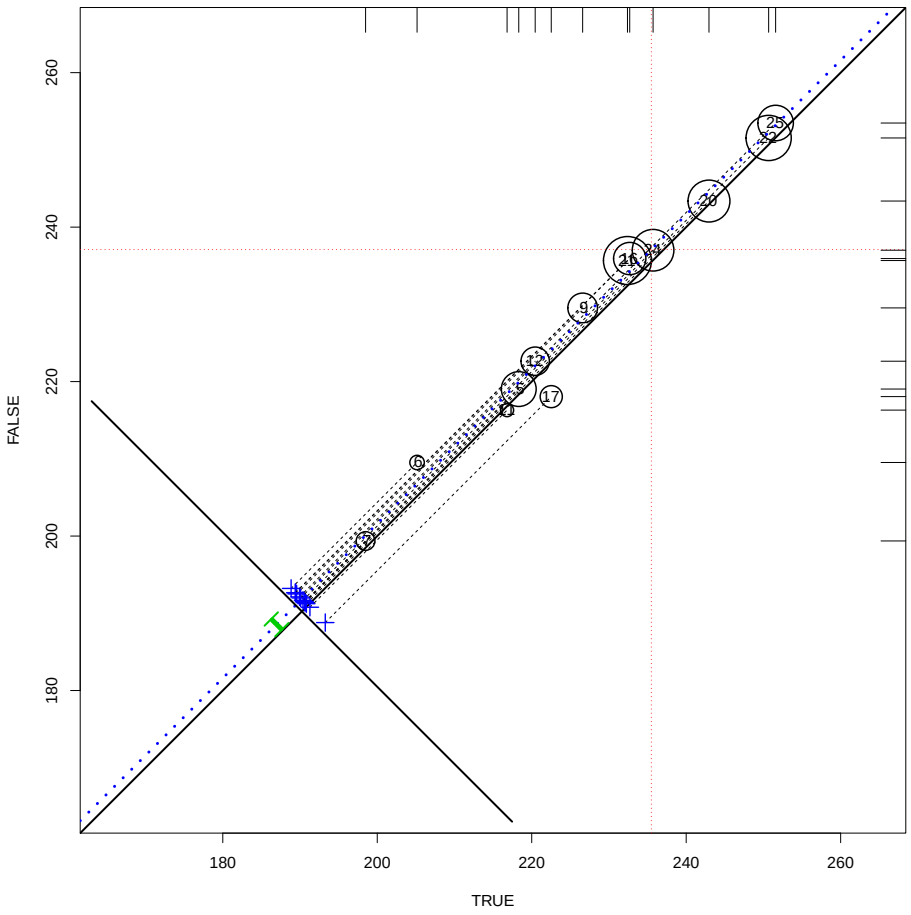


Figure 7: Propensity Score Assessment Plot for Classification Tree Stratification: Grade 4 Math

Table 5: Logistic Regression Stratification Results for Grade 4 math

	Public		Charter	
Strata	Mean	n	Mean	n
1	228.64	8702	225.26	158
2	237.61	8694	239.91	165
3	242.02	8637	239.31	222
4	246.25	8645	241.62	214
5	245.68	8607	248.17	253
6	247.34	8615	244.96	244
7	245.96	8566	242.53	293
8	234.39	8494	232.75	365
9	218.62	8246	217.66	613
10	224.42	8066	221.55	794

Table 6: Logistic Regression AIC Stratification Results for Grade 4 math

	Public		Charter	
Strata	Mean	n	Mean	n
1	229.81	8864	226.26	165
2	239.11	8630	241.24	162
3	245.14	11091	243.45	275
4	244.90	6095	242.28	160
5	244.90	8719	244.52	262
6	246.66	8483	246.44	250
7	245.45	8581	241.09	291
8	232.05	8507	231.18	349
9	218.70	8331	217.03	621
10	224.41	7971	222.01	786

Propensity Score Matching

The second class of propensity score method used is propensity score matching. In propensity score matching the goal is to match students from the two groups with small differences in their propensity scores. In large datasets, or when particular covariates are determined to be more important for adjusting selection bias, whether theoretically or otherwise, partial exact matching is done. In the context of this study partial exact matching is akin to implicitly adjusting for the multilevel nature of the data. That is, for this study, students are first matched exactly by state, gender, and ethnicity, then by propensity score. Propensity scores

Table 7: Classification Trees Stratification Results for Grade 4 math

	Public		Charter	
Strata	Mean	n	Mean	n
5	219.03	6783	218.32	492
6	209.53	731	205.16	78
7	199.37	1469	198.49	68
9	229.54	4489	226.60	376
11	216.31	595	216.81	38
12	222.64	3920	220.45	438
16	235.93	5918	232.69	110
17	218.06	2272	222.53	69
20	243.38	11290	242.95	260
21	235.67	15740	232.40	487
22	251.54	13499	250.68	478
24	237.01	11130	235.71	223
25	253.48	7436	251.58	204

from the full logistic regression model are used for matching.

The `Matchby` function in the `Matching` package (Sekhon, 2011) was used to find matches. First, propensity scores were estimated using the full logistic regression model. The `Matchby` algorithm first determines which students match exactly on state, gender, and ethnicity. Within those subgroups, students with the smallest standardized difference and less than 0.25 standard deviations, are returned. Three matched sets were produced (stated as charter-to-public): one-to-one, one-to-five, and one-to-ten. Matching was done without replacement.

Once matched pairs were determined, dependent sample t-tests were performed (Austin, 2011) to estimate average treatment effect and corresponding confidence intervals. Figures 12 and 13 and Table 8 at the end of the chapter provide the overall results. In general however, matching methods tend to estimate slightly larger treatment effects than both the stratification and multilevel models. And of additional note, the confidence intervals shrink as the ratio of treatment-to-control units increases.

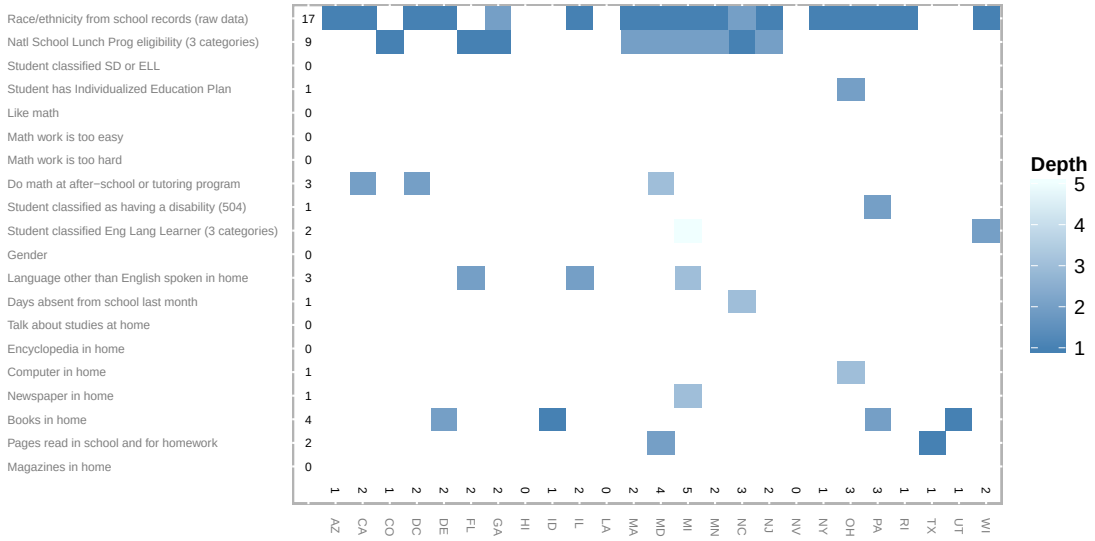


Figure 8: Multilevel PSA Covariate Heat Map for Classification Trees: Grade 4 Math

Multilevel Propensity Score Analysis

The final class of propensity score method utilized in multilevel propensity score analysis. This approach to PSA was developed for this dissertation and implemented in the `multilevelPSA` R package. The multilevel PSA approach makes explicit in both phase I and II the multilevel nature of the data dat,. in the case of this study, state. In principle the multilevel PSA approach is a conceptual combination of the partial exact matching and stratification. However, whereas partial exact matching utilizes propensity scores estimated from a single logistic regression model, the multilevel PSA algorithm will estimate separate propensity score models, using either logistic regression or classification trees, for each level two cluster (i.e. state). That is, the algorithm will perform m separate propensity score analyses using stratification where m is the number of states. This approach provides average treatment effects for each state as well as an overall, national, estimated treatment effect.

The same three methods of stratification described above will be used, full logistic regression model using all covariates, logistic regression model that

optimized Akaike Information Criterion (AIC), and classification trees. For the logistic regression models strata will be defined using quintiles of the propensity scores. One difficulty in interpreting results for multilevel PSA models is the relative importance of covariates for predicting treatment. Figure 77 is a covariate heat map that depicts each covariate on the y-axis and state on the x-axis. If a covariate is present in fitted classification tree for that state, the intersecting cell is shaded. The darkness of the color represents how far down the tree that covariate first appears. That is, the darkest color indicates that the covariate was used to split the tree at the root (or the first splitting covariate). This provides an opportunity to compare the relative importance of each covariate across states. The results for grade four math show that ethnicity is the strongest predictor of treatment having appeared in 17 of the trees with National School Lunch eligibility as the second. It should be noted that for the classification tree methods, strata with fewer than five students in either of the two groups were eliminated. Since quintiles were used for the logistic regression models, all students within those states are used. Table 5 provides the results within each each strata of each state including strata size.

Covariate Balance

Figure 9 is multilevel PSA counterpart to the covariate balance plot described above. Individual strata have been excluded for clarity since there substantially more strata. This figure shows that, in general, relatively good balance has been achieved since the adjusted absolute effect sizes are one, smaller or not substantially different than the unadjusted effect sizes, and two, all the adjusted effect sizes are smaller than 0.1. The remaining multilevel PSA covariate balance plots are provided in Appendix G. It should be noted that the classification tree methods, in general, provide much better balance than the logistic regression models. This is a limitation of estimating logistic regression models with smaller samples which have disproportional number of control-to-treatment students in the dependent variable. As such, interpreting the multilevel PSA logistic regression models in isolation is

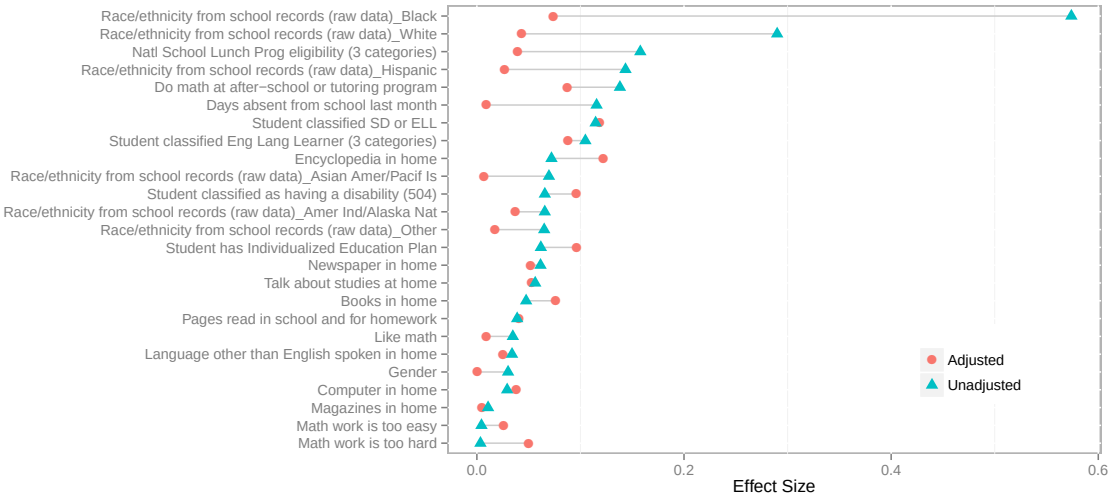


Figure 9: Multilevel PSA Covariate Balance Plot Classification Trees: Grade 4 Math

discouraged. However, this study follows the advise of Rosenbaum (2012) in that these are two of the nine methods used to estimate causal effects.

Visualizing Multilevel PSA

An important advantage of multilevel PSA is that average treatment effects can be estimated for each state and then aggregated to provide a national average treatment effect. A number of graphics have been developed to help interpret these results. Figure 10 is a multilevel PSA assessment plot for grade four math. This is an extension of the PSA assessment plots described above. The main panel (top-right) is a scatter plot where each point represents the overall adjusted score for each state (the point size is proportional to the number of students sampled in each state) with traditional public schools on the x-axis and charter schools on the y-axis. The panels to the bottom and left provide adjusted scores for each strata within each state as well as the overall state score for traditional public schools and charter schools, respectively. The overall national mean scores are represented by the blue lines. The tick marks on the line perpendicular to the unit line ($y = x$) represent the

distribution of differences for states. The dashed blue line[10] is the overall national mean difference and the green lines the 95% confidence interval. This figure depicts that there is now statistically significant difference nationally for grade four math using classification trees. Moreover, we see that there is minimal difference for most states since most of the points fall close to the unit line. Although there are some states that have a small positive effect size while others have a small negative effect.

Figure 11 provided a more detailed depiction of the differences depicted in Figure 10 as the tick marks on the line perpendicular to the unit line in the lower left corner of the plot. In Figure 11 the small grey points correspond to the difference within each strata. The blue points are the overall difference for each state; the point size corresponding to the number of students sampled; and 95% confidence intervals for each state in green. The overall adjusted national effect size and corresponding 95% confidence interval are represented by the vertical blue line and vertical green lines, respectively. Figures for grade four reading, grade eight math,and grade eight reading are provided in Appendix H.

[10]For this dataset the blue line almost complete overlaps the unit line but is present.

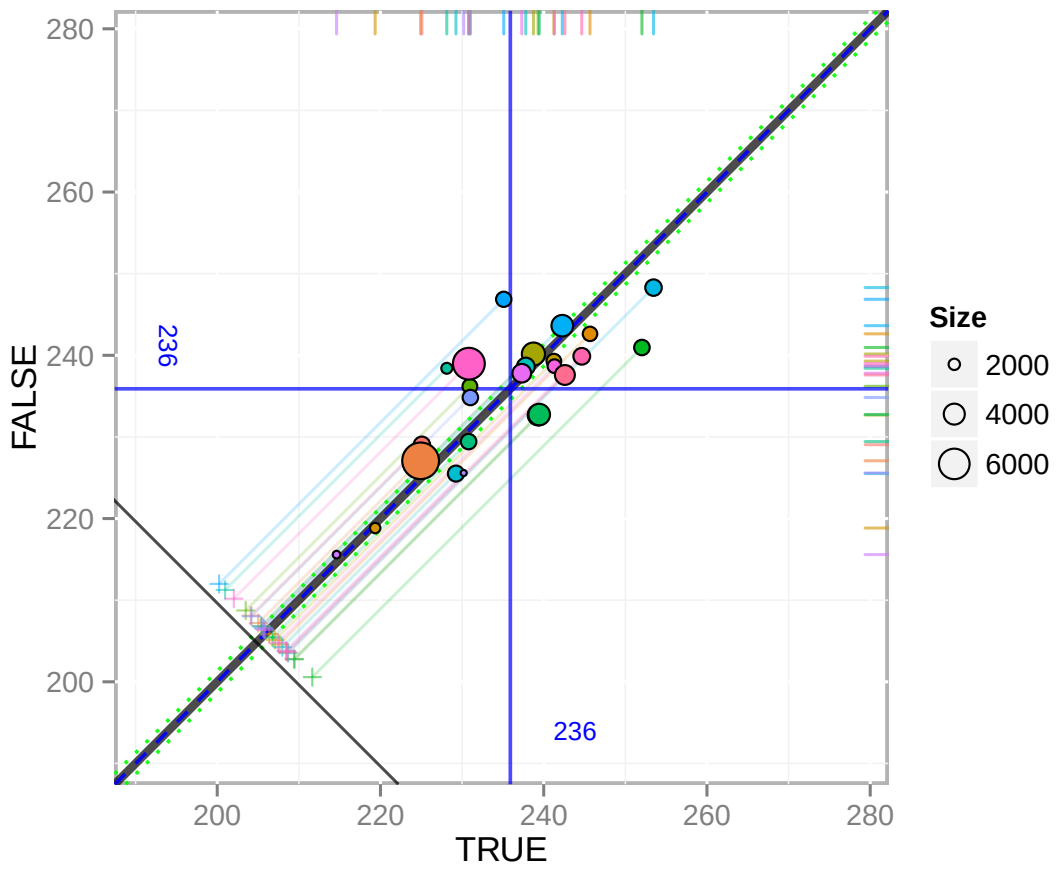


Figure 10: Multilevel PSA Assessment Plot Classification Trees: Grade 4 Math

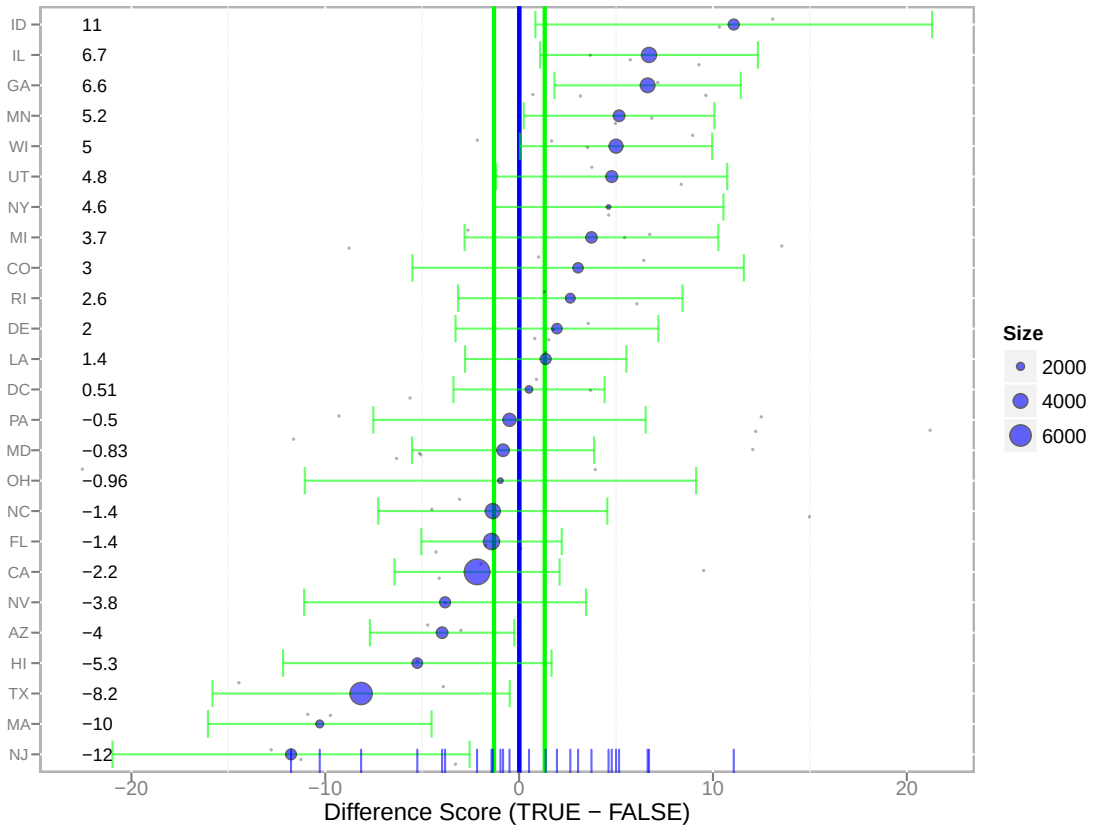


Figure 11: Multilevel PSA Difference Plot Classification Trees: Grade 4 Math

Summary and Overall Results

Up to this point in the chapter I have outlined the nine propensity score methods used for estimating treatment effects with grade four math. The corresponding tables and figures have been references in the appendices. In this section I will provide two figures and one table that summarize the 36 propensity score models estimated.

Figure 12 is a scatter plot of the overall national estimated treatment effects for all 36 PSA methods. The differences across subjects and grades is a result of different scales used for the assessment and therefore comparisons across subject and grade levels is not appropriate. The diameter of the circles in this figure are equal to the confidence interval so that circles that overlap the unit line indicate a non-significant difference. The horizontal and vertical lines (with numeric labels) represent the overall NAEP score for traditional public school students and charter school students, respectively. This figure shows that, in general, the scores for charter school students are in general higher when adjusted whereas the traditional public school scores are the same or lower. Regardless, in most cases the differences do not deviate substantially from the unit line indicating either no, or a small, difference is scores.

Figure 13 provides the overall national effect sizes for each PSA method within each grade and subject. Table 8 provide numeric results for this figure. This figure reveals a number of important conclusions. First, with regard to the effects of charter schools, there is some variety in effects across the different grades and subjects. In general, it appears charter schools either perform worse than or equal to traditional public schools in grade four. A few models within grade eight in both math and reading suggest small positive effects. However, even when there are statistically significant positive effects the maximum effect size is relatively small 0.1.

Secondly, Figure 13 reveal some trends in the behavior of the different propensity score methods. There appear to be fairly good consistency in the

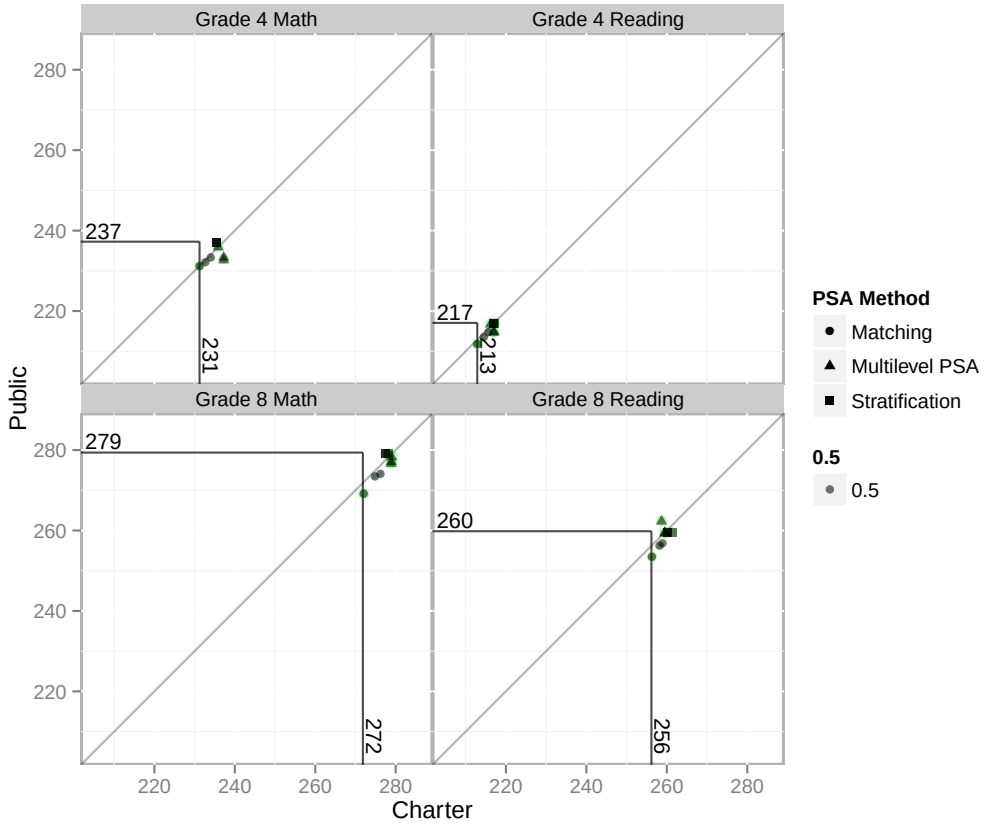


Figure 12: PSA Circle Plot of Adjusted Means

estimated effects within stratification and matching method although in general the matching methods provide larger effect size estimates. It should be noted that the matching methods, even with one-to-ten, use fewer than 40% of the available traditional public school students whereas the stratification methods use all traditional public school students. There is some variation in the estimated effect sizes for the multilevel models with the classification trees providing larger estimates. As noted above, this may be due, in part, to insufficient balance being achieved. This is likely a limitation of the logistic regression to provide stable estimates given one, the larger charter-to-public school student ratio and two, the smaller samples within each state. The following chapter will provide a discussion of the implications of these results.

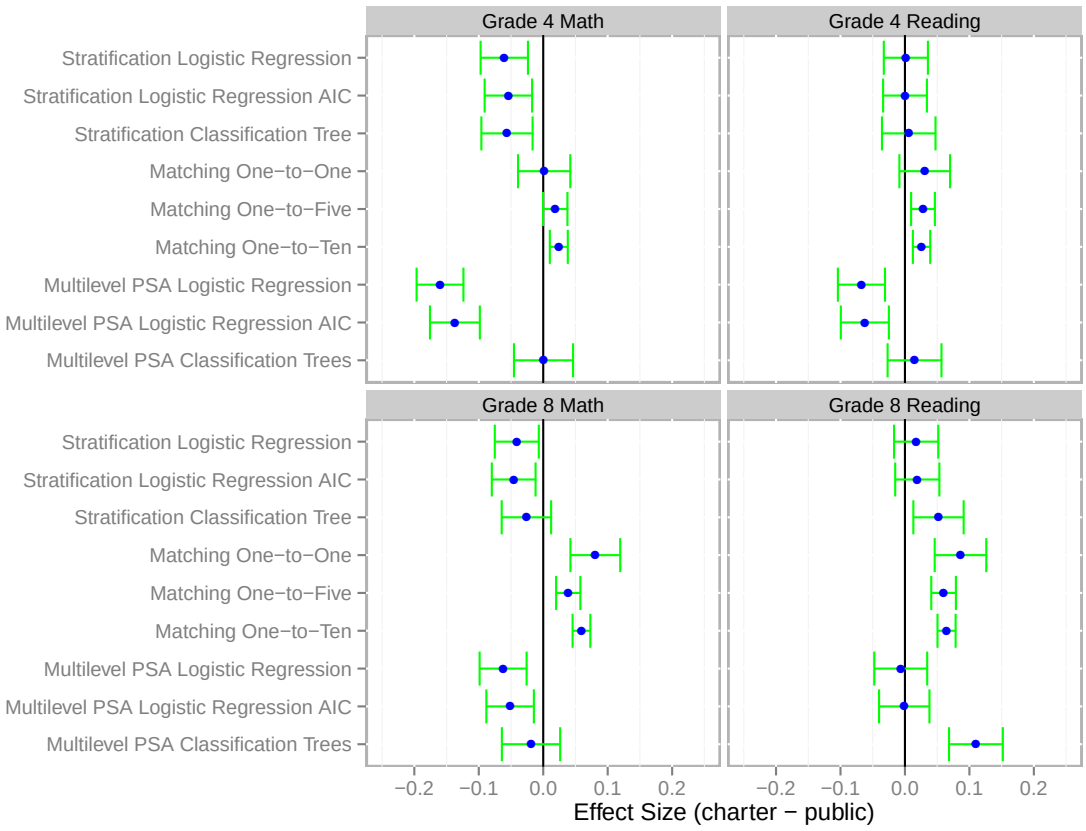


Figure 13: Overall Differences in Effect Size

Method	Charter	Public	ATE	95% CI	
			Grade 4 Math		
Stratification Logistic Regression	235.37	237.09	-0.06	-2.77	-0.68
Stratification Logistic Regression AIC	235.55	237.09	-0.05	-2.59	-0.49
Stratification Classification Tree	235.48	237.09	-0.06	-2.74	-0.47
Matching One-to-One	231.22	231.18	0.00	-1.12	1.20
Matching One-to-Five	232.67	232.14	0.02	-0.01	1.07
Matching One-to-Ten	234.02	233.33	0.02	0.29	1.09
Multilevel PSA Logistic Regression	237.24	232.67	-0.16	-3.53	-5.60
Multilevel PSA Logistic Regression AIC	237.24	233.33	-0.14	-2.80	-5.01
Multilevel PSA Classification Trees	235.90	235.91	0.00	1.32	-1.30
			Grade 4 Reading		
Stratification Logistic Regression	216.96	216.92	0.00	-1.13	1.23
Stratification Logistic Regression AIC	216.92	216.92	0.00	-1.17	1.17
Stratification Classification Tree	217.12	216.92	0.01	-1.23	1.63
Matching One-to-One	212.88	211.82	0.03	-0.30	2.41
Matching One-to-Five	214.51	213.54	0.03	0.33	1.60
Matching One-to-Ten	215.63	214.75	0.03	0.42	1.35
Multilevel PSA Logistic Regression	217.01	214.69	-0.07	-1.07	-3.58
Multilevel PSA Logistic Regression AIC	216.99	214.85	-0.06	-0.86	-3.42
Multilevel PSA Classification Trees	216.19	216.71	0.01	1.95	-0.93
			Grade 8 Math		
Stratification Logistic Regression	277.58	279.05	-0.04	-2.69	-0.25
Stratification Logistic Regression AIC	277.41	279.06	-0.05	-2.86	-0.43
Stratification Classification Tree	278.11	279.04	-0.03	-2.31	0.44
Matching One-to-One	272.05	269.16	0.08	1.51	4.28
Matching One-to-Five	274.85	273.46	0.04	0.72	2.06
Matching One-to-Ten	276.21	274.08	0.06	1.63	2.62
Multilevel PSA Logistic Regression	278.86	276.63	-0.06	-0.92	-3.54
Multilevel PSA Logistic Regression AIC	278.95	277.11	-0.05	-0.52	-3.16
Multilevel PSA Classification Trees	278.98	278.30	-0.02	0.94	-2.30
			Grade 8 Reading		
Stratification Logistic Regression	260.20	259.63	0.02	-0.55	1.69
Stratification Logistic Regression AIC	260.25	259.63	0.02	-0.49	1.74
Stratification Classification Tree	261.30	259.60	0.05	0.43	2.97
Matching One-to-One	256.29	253.48	0.09	1.51	4.12
Matching One-to-Five	258.19	256.24	0.06	1.33	2.58
Matching One-to-Ten	258.93	256.82	0.06	1.65	2.57
Multilevel PSA Logistic Regression	259.51	259.30	-0.01	1.13	-1.55
Multilevel PSA Logistic Regression AIC	259.52	259.49	-0.00	1.24	-1.31
Multilevel PSA Classification Trees	258.70	262.29	0.11	4.96	2.23

Table 8: Summary of Overall Propensity Score Results

CHAPTER 5: DISCUSSION

This study aims to make two major contributions: first, develop a new method of propensity score analysis for multilevel data, and two, address the question of the effectiveness of charter schools from a state and national perspective. This chapter will provide some concluding remarks and conclusions as well as point some limitations of this study.

Multilevel Propensity Score Analysis

The development of the `multilevelPSA` R package for estimating and visualizing propensity score models of multilevel data provides important insight into the implications of what traditionally would have been one of many covariates. The results suggest that this method performs and provide effect size estimates consistent with other propensity score methods. However, a key advantage to using this new method include an explicit adjustment of the multilevel nature of some data as well as being able to understand the implication of clustering when comparing the outcome of interest. As researchers work with larger and larger datasets, the advancements in data visualization should not be understated. The data visualizations developed for this dissertation provide important insight into data analysis at all stages.

Propensity score methods have shown to be effective for estimating treatment effects with relatively small samples. However, as observed in chapter four, estimating multilevel PSA models require larger samples given the need to stratify within clusters. Although NCES began oversampling charter schools in 2005, the fact that ratio of charter to traditional public schools/students is so large result in model specification problems, especially with logistic regression. This has been alleviated to some extent by removing traditional public school students who attend a school further than five miles from a charter school.

More specifically with regard to propensity score ranges, the range tends to

shrink as the ratio of treatment-to-control increases. Figure 14 depicts the range and distribution of propensity scores (using logistic regression) with varying treatment-to-control ratios. The data used to create this figure is simulated and available in Appendix K. The `psrange` and `plot.psrange` functions are included in the `multilevelPSA` R package. Propensity scores are estimated with a single covariate where the mean for the treatment and control are 0.6 and 0.4, respectively. The standard deviation for both are 0.4. There are 100 treatment units and 1,000 control units simulated. The goal in choosing these means and standard deviations is to have some separation between treatment and control. Each row in the figure represents the percentage of control units sampled before estimating the propensity scores starting with 100% (i.e. all 1,000 control units) to 10% (100 of the control units). As the figure shows, as the ratio decreases to where there are equal treatment and control units, the range of the propensity scores becomes more normal. To calculate the ranges each sampling step is bootstrapped so the green bar and black points represent each of the 20 bootstrap staples taken. The bars then represent the mean of minimum and mean of the maximum for each step.

The Display of Multilevel Results

In the development of the `multilevelPSA`, as well as all the analysis in the dissertation, two overarching principal decisions were made with regard to how results are displayed, specifically the lack of p-values and an emphasis on visualizations over tabular output. Both these issues have received substantial attention and debate over the last several decades. And although there is no clear consensus on "best-practices," I contend that given the nature of propensity score analysis and observational studies, simple null hypotheses reported as either statistically significant or not in tables with `p`-values does a disservice to the results. The use of graphics with confidence intervals provide context as well as magnitudes of differences.

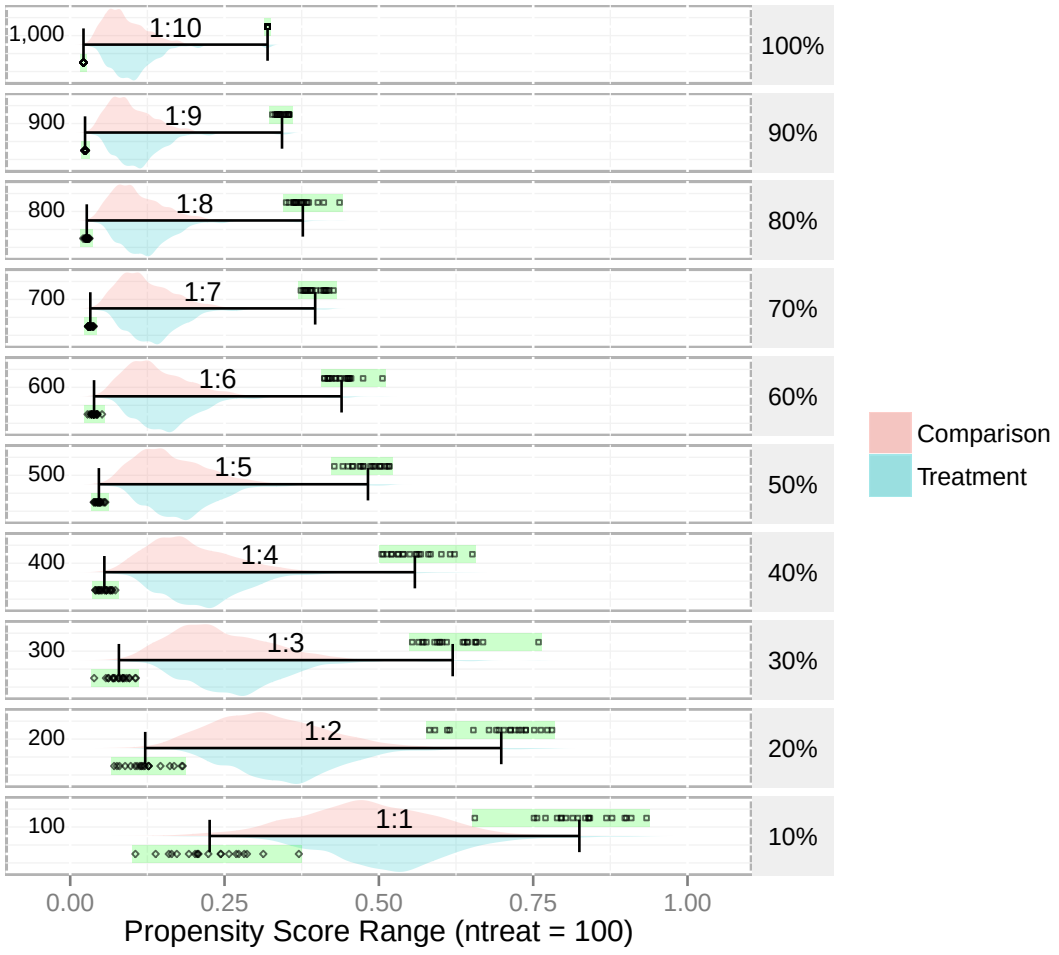


Figure 14: Propensity Score Ranges for Varying Treatment-to-Control Ratios

The practice of significance testing[11] dates to the early work of Fisher (1925). Gigerenzer (2004) describes the current practice in peer-reviewed research journals as "the null ritual" that involve three steps:

1. Define a null hypothesis where the researcher is testing that there is no mean difference. Also, don't specify any predictions or alternative hypotheses.

2. Use a p-value of .05 for rejecting the null hypothesis and report your p-value using a range (i.e. $p < .05$, $p < .01$, or $p < .001$).

[11] Here, I use the phrases significance testing, null hypothesis testing, and p-values to represent the same statistical practice and are generally interchangeable.

3. Always perform this procedure.

In 1996 the American Psychological Association (APA) brought the debate regarding the use of significance testing to the forefront by entertaining a ban in the journals it publishes (Shrout, 1997; Hunter, 1997; Harris, 1997; Abelson, 1997; Scarr, 1997; Estes, 1997b). Although a ban was not instituted, APA now recommends the reporting of exact p-values, confidence intervals, and effect sizes.

What is the issue with p-values? First, the practice of significance testing as represented in the social sciences for nearly the last century reduce research question to a dichotomous outcome. Rarely can a study be reduced to simple yes/no answer, especially in the social sciences. Moreover, the use of $p < .05$ is entirely arbitrary and the difference between significant and non-significant results is itself not significant (Gelman & Stern, 2006). However, perhaps more damning is the likelihood of committing Type II errors (Bakan, 1966; Carver, 1978; Cohen, 1994; Henkel & Morrison, 1970; Rozenboom, 1960; Schmidt, 1996). In a study by Sedlmeier and Gigerenzer (1989) that examined all published articles in 1984 in the *Journal of Abnormal Psychology* found that the error rate was 60%. That is to say that researchers would have done better to flip a coin!

Lastly, given the relationship between p-values and sample size, it would be expected that with n >100,000, as in this study, most differences would be statistically significant. Even in one-to-one matched analysis where $n \approx 3,000$, we would expect p <0.05 for even small differences. The formula for calculating t for a dependent sample paired t-test is:

$$t = \frac{\bar{X}_D - \mu_0}{S_D/\sqrt{n}}$$

Where X_D is the mean difference, μ_0 is non-zero for testing differences other than zero, S_D is the standard deviation of the differences, and n is the sample size. Using the approximate sample standard deviation of 40 from grade 8 reading results, a mean difference 2 (representing a small effect size of 0.06 by most standards), and

$n = 3000$, we get $t = 2.738$ and $p = 0.003$. For the one-to-one paired analysis with a very small effect size the power estimate Cohen (1977) is 0.64. However, the one-to-two matched analysis increases the power to a very acceptable 0.90. The results of this exercise is to demonstrate that relying on p-values to make decisions is a *fool's errand*. Instead, we should take Scarr's suggestion that "better uses of statstics would focus on the magnitude of effects and error estimates" (Scarr, 1997).

The use of graphics have made substantial advancements in the twentieth century with seminal works by Tukey (Cleveland, 1988), Tufte (2001), Cleveland (1993, 1994), and Chambers, Cleveland, Kleiner, and Tukey (1983). The implementations utilized in this dissertation are based upon Wilkinson's (2005) *grammar of graphics* as implemented in R using the `ggplot2` package (Wickham, 2009). Wherever possible, confidence intervals are used to show the magnitude of the differences (c.f. Cumming, 2012; Estes, 1997a). Although the use of graphics are frequently taught in statistics courses, they are often omitted from journal publications (Gelman, Pasarica, & Dodhia, 2002) and relegated to diagnostic purposes (Gelman, 2011). The graphics presented here provide important insight into the nature and magnitude of the differences between charter and traditional public schools. The multilevel assessment plots (see Figure 6 and Appendix H) show the distribution scores (charter, traditional public, and differences) across multiple dimensions simultaneously. Perhaps we would see from a table that the differences are small within states with few exceptions. But what would be lost is that the range of scores across states for charter and traditional public school students separately is relatively wide. Similarly, for the multilevel PSA difference plots (see Figure 11 and Appendix H) the graphic provides immediate evidence of the nature of the differences. Also, by providing confidence intervals, we can express results vis--vis the graphic similar to the traditional p-value in a summary table.

Differences Between Charter and Traditional Public Schools

Given the substantial difference in sample n's for charter and public schools (i.e. there are as much as three to four orders of magnitude more public schools students available in the NAEP data sets), it is expected that there would be public school students who would not have a counterpart from the charter school group. However, the relatively high percentage of public schools students who do not have a charter school counterpart (as much as 35%) suggest that there may be imbalance between the two groups as a whole. That is, although reasonable balance was achieved with regard to the individual strata where comparisons are made, the overall sample imbalance, as evidenced by the unmatched public school students, suggests that public schools serve a more heterogeneous population.

Moreover, the methodological issues described above may, in part, be evidence of any underlying fundamental difference in charter school and traditional public school populations. Figure 13 in chapter four reveals that the methods with higher effect sizes are the matching methods and classification trees. These models, especially the matching methods, often eliminate a large proportion of tradition public school students. Perhaps there may be a small positive effect of charter schools for a very specific type of student, but there also appears to be a large number of students who appear better served, or at least equivalently served, by traditional public schools. This needs to be considered when weighing the cost and extent of charter schools.

It should also be noted that significant limitation of this study, and many like it, is that it only examines a small subset of a students educational experiences. That is, NAEP as well as CREDO only examine student performance in math and reading. This leaves out all other, and arguably equally important, subjects. These studies are not alone in overemphasizing these subjects. The Common Core State Standards which being implemented in the majority of states and is a cornerstone of

the *Race to the Top* initiative, only defines standards and curriculum for mathematics and English Language Arts. The emphasis on only these two subjects has resulted in many schools reducing or eliminating the arts and other subjects to spend more school time preparing students in these subjects (c.f. Ravitch, 2013).

What is the Relationship Between Charter School Performance and Charter School Laws?

The National Alliance for Public Charter Schools (NAPCS) publishes rankings of state charter school laws (Ziebarth, 2012). They argue that some the performance of charter schools in some states may be hindered by state law. Figure 15 is a slepegraph (Tufte, 2001) that compares the rankings given to each state by NAPCS with that state's ranking using stratification with classification trees[12]. Although top rated states by NAPCS had larger differences in NAEP scores, there is considerable crossing of lines between the two rankings. This visual cue indicates that the rankings are not very correlated. It should also be noted that not all states that have charter laws had sufficient sample size in NAEP to be included.

In summary, these results are consistent with the wide body of research on charter schools. Namely, some charter schools better than their traditional public school counterpart, while others perform worse. However, in aggregate, the average difference is very small. Ray Budde originally envisioned charter schools as a way for teachers, administrators, and communities to experiment with the goal of finding better ways to teach students. But with *No Child Left Behind* and *Race to the Top*, among other initiatives from private for-profit and not-for-profit organizations, charter schools are often offered as a wholesale replacement for traditional public schools. However, these results along with the other national studies examining the differences between charter and traditional public schools suggest that charter schools do not provide, in aggregate, substantial benefit over their traditional public school counterparts.

[12] A slopegraph for only the stratification with classification trees is provided since 1) it resulted in better balance as compared to the other methods, and 2) the results are consistent regardless of method used.

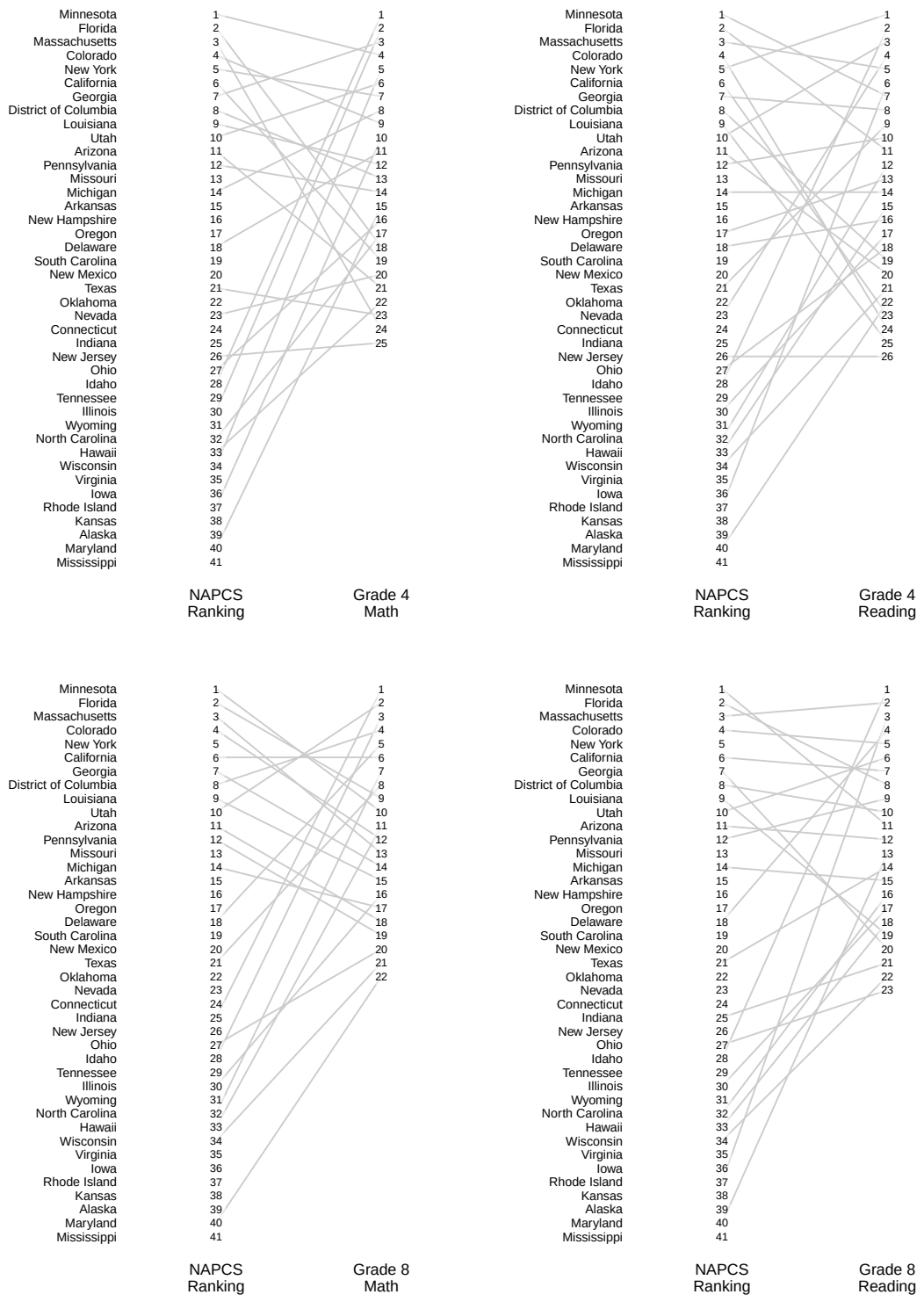


Figure 15: Comparison of 2012 National Alliance for Public Charter Schools (NAPCS) State Charter School Law Rankings and NAEP Charter School Rankings

REFERENCES

Abadie, A., Diamond, A., & Hainmueller, J. (2007). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *NBER Working Paper Series*, *w12831*. Retrieved from `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=958483`

Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science*, *8*(1), 12–15.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716-723.

Allen, J., Consolettie, A., & Kerwin, K. (2009). *The accountability report: Charter schools.* The Center for Education Reform.

Arpino, B., & Mealli, F. (2008). *The specification of the propensity score in multilevel observational studies.* Munich Personal RePEc Archive. Retrieved from `http://mpra.ub.uni-muenchen.de/17407`

Austin, P. C. (2011). Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched smaples. *Statistical in Medicine*, *30*, 1292-1301.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*, 423–437.

Betts, J. R., & Hill, P. T. (2006). *Key issues in studying charter schools and achievement: A review and suggestions for national guidelines.* National Charter School Research Project, Center on Reinventing Public Education, University of Washington Bothell.

Betts, J. R., & Tang, Y. E. (2008). *Value-added and experimental studies of the effect of charter schools on student achievement.* National Charter School Research Project, Center on Reinventing Public Education, University of Washington Bothell.

Braun, H., Jenkins, F., & Grigg, W. (2006a). *A closer look at charter schools using hierarchical linear modeling.* U.S. Government Printing Office.

Braun, H., Jenkins, F., & Grigg, W. (2006b). *Comparing private schools and public schools using hierarchical linearl modeling (nces 2006-461).* Washington, DC: U.S. Department of Education, National Center for Educational Studies, Institute of Education Sciences.

Bryer, J. (2011). multilevelpsa: Multilevel propensity score analysis [Computer software manual]. Retrieved from http://multilevelpsa.r-forge.r-project.org (R package version 1.0)

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage Publications.

Budde, R. (1988). *Education by charter: Restructuring school districts*. The Regional Laboratory for Education Improvement.

Carnoy, M., Jacobsen, R., Mishel, L., & Rothstein, R. (2005). *The charter school dust-up: Examining the evidence on enrollment and achievement*. Teacher College Press & Economic Policy Institute.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*, 378–399.

Center for Education Reform. (2008). *Annual survey of america's charter schools*. Retrieved November 29, 2009, from http://edreform.com/wp-content/uploads/2013/03/CER_charter_survey_2008.pdf

Center for Education Reform. (2010). *National charter school & enrollment statistics*. Retrieved from http://www.edreform.com/download/CER_Charter_Survey_2010.pdf

Center for Research on Education Outcomes. (2009). *Multiple choice: Charter school performance in 16 states*. Stanford University.

Center for Research on Education Outcomes. (2013). *National charter school study*. Stanford University.

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical methods for data analysis*. Belmont.

Cleveland, W. S. (1988). *The collected works of john w. tukey, volume v graphics*. Wadsworth and Brooks.

Cleveland, W. S. (1993). *Visualizing data*. AT&T Bell Laboratories.

Cleveland, W. S. (1994). *The elements of graphing data (rev. ed.)*. AT&T Bell Laboratories.

Cohen, J. (1977). *Statistical power for the behavioral sciences (2nd ed.)*. Academic Press.

Cohen, J. (1994). The earth is round (p ¡ .05). *American Psychologist*, *49*, 997–1003.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, *27*(4).

Cullen, J. B., Jacob, B. A., & Levitt, S. D. (2005). The impact of school choice onstudent outcomes: An analysis of the Chicago Public Schools. *Journal of Public Economics*, *89*, 729-760.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94*(448), 1053–1062.

Estes, W. K. (1997a). On the communication of information by displays of standard errors and confidence intervals. *Psychological Bulletin & Review*, *4*(3), 330–341.

Estes, W. K. (1997b). Significance testing in psychological research: Some persisting issues. *Psychological Science*, *8*(1), 18–20.

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver and Boyd.

Foundations for the Future Charter Academy. (2007, May). *Charter document.* Retrieved from `http://www.ffca-calgary.com/downloads/Charter.pdf`

Gelman, A. (2011). Why tables are really much better than graphs. *Journal of Computational and Graphical Statistics*, *20*(1), 3–7.

Gelman, A., & Hill, J. (2006). *Data analysis using regressino and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Gelman, A., Pasarica, C., & Dodhia, R. (2002). Let's practice what we preach: Turning tables into graphs. *The American Statistician*, *56*(2), 121–130.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *American Statistician*, *6-*(4), 328–331.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*, 587–606.

Harris, R. J. (1997). Significance tests have their place. *Psychological Science*, *8*(1), 8–11.

Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1997). *Characterizing selection bias using experimental data.* Retrieved from `http://athens.src.uchicago.edu/jenni/dvmaster/FILES/matchingf.pdf`

Helmreich, J. E., & Pruzek, R. M. (2009). PSAgraphics: An R package to support propensity score analysis. *Journal of Statistical Software*, *29*(6).

Henkel, R., & Morrison, D. (1970). *The significance test controversy*. Butterworth.

Herbst, J. (2006). *School choice and school governance: A historical study of the United States and Germany.* Palgrave Macmillan.

Hofstede, G., & Hofstede, G. J. (2004). *Cultures and organizations: Software of the mind* (2nd ed.). New York, NY: McGraw Hill.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945-960.

Hubbard, L., & Kulkarni, R. (2009). Charter schools: Learning from the past, planning for the future. *Journal of Educational Change*, *10*, 173-189.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, *8*(1), 3–7.

Kolderie, T. (2005). Ray budde adn the origins of the charter school.

Lander, M. (2001). *School chice, Kiwi-style: When New Zealand abolished school boards.* (Tech. Rep.). Frontier Centre for Public Policy.

Larrañaga, O. (2004). *Competencia y participación privada: la experiencia chilena en educación.* (Tech. Rep.). Estudios Publicos.

Loveless, T. (2013). *Charter school study: Much ado about tiny differences.* Retrieved from `http://www.brookings.edu/blogs/brown-center-chalkboard/posts/2013/07/03-charter-schools-loveless`

Maccall, W. (1847). *The elements of individualism.* London: John Chapman.

National Alliance of Public Charter Schools. (2009). *Charter school achievement: What we know.* Retrieved from `http://www.publiccharters.org/What+We+Know+5`

National Assessment Governing Board. (2006a). *Mathematics framework for the 2007 national assessment of educational progress.* Washington, DC: U.S. Department of Education.

National Assessment Governing Board. (2006b). *Reading framework for the 2007 national assessment of educational progress.* Washington, DC: U.S. Department of Education.

National Center for Educational Statistics. (2009). *Common core of data.* Retrieved from `http://nces.ed.gov/ccd`

Nelson, F. H., Rosenberg, B., & Meter, N. V. (2004). *Charter school achievement on the 2003 national assessment of educational progress.* Washington, DC: American Federation of Teachers. Retrieved from `http://www.aft.org/pubs-reports/downloads/teachers/NAEPCharterSchoolReport.pdf`

Northwest Evaluation Association. (2009). *Why is the growth research database significant.*

Retrieved from http://www.nwea.org/support/details.aspx?content=1053

Organisation for Economic Co-Operation and Development. (2009). Programme of international student assessment [Computer software manual]. Retrieved from http://www.oecd.org/pisa/

Pearl, J. (2009). *Causality: Models, reasoning and inference (2nd ed.)*. Cambridge University Press.

R Development Core Team. (2008). *[computer software]. R: A language and environment for statistical computing. r foundation for statistical computing.* Vienna, Austria.

Raudenbush, S. W., Hong, G., & Rowan, B. (2003). *Studying the causal effects of instruction with application to primary-school mathematics. invited talk at the research seminar ii: Instructional and performance consequences of high poverty schooling.*

Ravitch, D. (2013). *Reign of error.* Random House, Inc.

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

Rosenbaum, P. R. (2012). Testing one hypothesis twice in observational studies. *Biometrika*, *99*, 763-774.

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

Rozenboom, W. (1960). The fallacy of the null-hypothesis significance test. *lotical Bulletin*, *57*, 316–428.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*, 473-489.

Scarr, S. (1997). Rules of evidence: A larger context for the statistical debate. *Psychological Science*, *8*(1), 16–17.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge chology: Implications for training of researchers. *Psychological Methods*, *129*.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs.* American Educational Research Association.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*(2), 309–316.

Sekhon, J. S. (2011, 6 14). Multivariate and propensity score matching software with automated balance optimization: The matching package for r. *Journal of Statistical Software*, *42*(7), 1–52. Retrieved from `http://www.jstatsoft.org/v42/i07`

Shadish, W. R. (2013). Propensity score analysis: promise, reality and irrational exuberance. *Journal of Experimental Criminology*, *9*(2), 128–144.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? a randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, *103*(484), 1334–1343.

Shrout, P. E. (1997). Should significance tests be banned? *Psychological Science*, *8*(1), 1–2.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*, 1-21.

Stuart, E. A., & Rubin, D. B. (2008). Best practices in quantitative methods. In J. Osborne (Ed.), (p. 155-176). Sage Publications.

Swart, K. W. (1962). "Individualism" in the mid-nineteenth century. *Journal of the History of Ideas*, *23*(1), 77-90.

Teske, P., & Schneider, M. (2001). What research can tell policymakers about school choice. *Journal of Policy Analysis and Management*, *20*(4), 609-631.

The National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform.

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

Tufte, E. (2001). *The visual display of quantitative information.* Graphics Press LLC.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*(3), 1–67. Retrieved from `http://www.jstatsoft.org/v45/i03/`

van Buuren, S., & Groothuis-Oudshoorn, K. (n.d.). MICE: Multivariate imputation by chained equations in r. *Journal of Statistical Software*.

Vanourek, G., Manno, B., Finn, C., & Bierlein, L. (1998). Charter schools. In B. Hassel & P. Peterson (Eds.), *Learning from school choice* (p. 187-211). Washington, DC: Brookings Institution.

Wells, A. S. (Ed.). (2002). *Where charter school policy fails: The problems of accountability and*

equity. New York, NY: Teachers College Press.

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. Springer.

Wilkinson, L. (2005). *The grammar of graphics (2nd ed)*. Springer.

Ziebarth, T. (2012). *Measuring up to the model: A ranking of state charter school laws* (3rd ed.). Washington D.C.: National Alliance for Public Charter Schools.

Appendix A

Charter Schools & Student Enrollment by State

Table 9: Charter Schools & Student Enrollment by State

State	Law Enacted	Totals for Charter Schools[b]			NAEP Students	
		Operating	Closed	Students	Charters	Publics
Alabama[a]		0	0	0	0	2759
Alaska	1995	26	5	5,198	69	2517
Arizona	1994	510	96	119,903	99	2674
Arkansas	1995	25	6	6,750	30	2407
California	1992	802	103	316,468	417	7803
Colorado	1993	151	10	54,497	108	2598
Connecticut	1996	21	5	3,932	0	2531
Delaware	1995	21	2	8,740	180	2641
Washington DC	1996	93	16	25,385	652	1336
Florida	1996	382	82	108,382	175	3876
Georgia	1993	83	5	40,807	64	3465
Hawaii	1994	32	0	7,317	132	2605
Idaho	1998	32	1	10,492	59	2784
Illinois	1996	74	8	27,683	33	4015
Indiana	2001	50	2	12,631	11	2720
Iowa	2002	10	0	1,462	0	2839
Kansas	1994	40	10	3,361	17	2726
Kentucky[a]		0	0	0	0	2696
Louisiana	1995	66	10	23,634	97	2264
Maine[a]		0	0	0	0	2658
Maryland	2003	34	2	7,301	6	2825
Massachusetts	1993	64	6	23,905	56	3667
Michigan	1993	250	27	94,092	134	2480
Minnesota	1991	159	29	28,371	16	2875
Mississippi	1997	1	0	367	0	2613
Missouri	1998	39	5	13,125	38	2771
Montana[a]		0	0	0	0	2581
Nebraska[a]		0	0	0	0	2688
Nevada	1997	26	7	7,295	0	2662
New Hampshire	1995	11	2	1,212	0	2803
New Jersey	1996	64	19	17,986	0	2813
New Mexico	1993	70	3	11,426	54	2722
New York	1998	118	10	32,602	16	3745
North Carolina	1996	103	32	30,445	72	4090
North Dakota[a]		0	0	0	0	2307
Ohio	1997	293	48	94,171	45	3746
Oklahoma	1999	14	1	4,770	0	2612
Oregon	1999	93	8	13,612	41	2626
Pennsylvania	1997	133	12	61,823	64	2709
Rhode Island	1995	11	0	2,894	30	2621
South Carolina	1996	36	10	8,705	16	2697
South Dakota[a]		0	0	0	0	2889

Charter Schools & Student Enrollment by State (cont.)

State	Law Enacted	Totals for Charter Schools[b]			NAEP Students	
		Operating	Closed	Students	Charters	Publics
Tennessee	2002	14	1	2,585	54	2815
Texas	1995	331	33	108,541	199	7070
Utah	1998	68	1	23,233	38	2722
Vermont[a]		0	0	0	0	2003
Virginia	1998	4	3	275	0	2848
Washington[a]		0	0	0	0	2968
West Virginia[a]		0	0	0	0	2831
Wisconsin	1993	221	37	41,799	114	2592
Wyoming	1995	3	0	244	0	1897
Total		4,578	657	1,407,421	3,164	156,963

[a]State currently does not have a charter school law.
[b]Source: Center for Education Reform (2010)

Appendix B

Descriptive Statistics

Table 10: Grade 4 Math Descriptive Statistics

	Traditional		Charter	
Race/ethnicity from school records (raw data)				
White	90268	57%	1202	33%
Black	27565	17%	1546	43%
Hispanic	27927	18%	642	18%
Asian Amer/Pacif Is	7657	5%	172	5%
Amer Ind/Alaska Nat	3753	2%	35	1%
Other	2168	1%	28	1%
Unknown	0	0%	0	0%
Natl School Lunch Prog eligibility (3 categories)				
Eligible	79160	50%	2074	57%
Not eligible	79273	50%	1381	38%
Info not available	905	1%	170	5%
Unknown	0	0%	0	0%
Student has Individualized Education Plan				
Yes, IEP	17871	11%	332	9%
Yes, 504 plan	1414	1%	23	1%
Yes, 504 in process	0	0%	0	0%
Not IEP	140022	88%	3270	90%
Omitted	0	0%	0	0%
Unknown	31	0%	0	0%
Student classified Eng Lang Learner (3 categories)				
Yes	13002	8%	275	8%
No	143127	90%	3273	90%
Formerly ELL	3174	2%	77	2%
Omitted	0	0%	0	0%
Unknown	35	0%	0	0%
Gender				
Male	81536	51%	1796	50%
Female	77802	49%	1829	50%
Unknown	0	0%	0	0%
Student classified as having a disability (504)				
Student with disabi	19285	12%	355	10%
Not student with di	140022	88%	3270	90%
Omitted	31	0%	0	0%
Unknown	0	0%	0	0%
Student classified SD or ELL				
Student with disabi	17847	11%	328	9%
English language le	11564	7%	248	7%
Both SD and ELL	1438	1%	27	1%
Neither SD nor ELL	128441	81%	3022	83%
Unknown	48	0%	0	0%
Newspaper in home				

continued on next page...

...continued from previous page

	Charter		Traditional	
Yes	44894	28%	966	27%
No	55957	35%	1334	37%
I Don't Know	55462	35%	1210	33%
Omitted	3004	2%	115	3%
Multiple	21	0%	0	0%
Unknown	0	0%	0	0%
Magazines in home				
Yes	89988	56%	1998	55%
No	38593	24%	877	24%
I Don't Know	27543	17%	627	17%
Omitted	3190	2%	123	3%
Multiple	24	0%	0	0%
Unknown	0	0%	0	0%
Books in home				
0-10 books	19625	12%	423	12%
11-25 books	33693	21%	825	23%
26-100 books	52311	33%	1079	30%
More than 100 books	50511	32%	1181	33%
Omitted	3159	2%	117	3%
Multiple	39	0%	0	0%
Unknown	0	0%	0	0%
Computer in home				
Yes	136033	85%	3078	85%
No	19502	12%	413	11%
Omitted	3787	2%	134	4%
Multiple	16	0%	0	0%
Unknown	0	0%	0	0%
Encyclopedia in home				
Yes	80440	50%	1874	52%
No	25501	16%	514	14%
I Don't Know	50222	32%	1120	31%
Omitted	3146	2%	114	3%
Multiple	29	0%	3	0%
Unknown	0	0%	0	0%
Pages read in school and for homework				
5 or fewer	33661	21%	819	23%
6-10	27785	17%	618	17%
11-15	20828	13%	442	12%
16-20	22687	14%	488	13%
More than 20	51046	32%	1134	31%
Omitted	3259	2%	124	3%
Multiple	72	0%	0	0%
Unknown	0	0%	0	0%
Talk about studies at home				
Never or hardly eve	29003	18%	578	16%
Every few weeks	21264	13%	468	13%
About once a week	18741	12%	400	11%
2-3 times a week	31451	20%	680	19%

continued on next page...

...continued from previous page

	Charter		Traditional	
Every day	55569	35%	1377	38%
Omitted	3247	2%	121	3%
Multiple	63	0%	1	0%
Unknown	0	0%	0	0%
Days absent from school last month				
None	79833	50%	1622	45%
1-2 days	46548	29%	1111	31%
3-4 days	18267	11%	434	12%
5-10 days	7351	5%	207	6%
More than 10 days	4078	3%	130	4%
Omitted	3207	2%	120	3%
Multiple	54	0%	1	0%
Unknown	0	0%	0	0%
Language other than English spoken in home				
Never	85236	53%	1679	46%
Once in a while	33507	21%	833	23%
Half the time	11284	7%	297	8%
All or most of time	26049	16%	695	19%
Omitted	3214	2%	121	3%
Multiple	48	0%	0	0%
Unknown	0	0%	0	0%
Do math at after-school or tutoring program				
Yes	53627	34%	1532	42%
No	101907	64%	1955	54%
Omitted	3780	2%	137	4%
Multiple	24	0%	1	0%
Unknown	0	0%	0	0%
Math work is too hard				
Never or hardly eve	46369	29%	1028	28%
Sometimes	87164	55%	1964	54%
Often	14112	9%	305	8%
Always or almost	7374	5%	177	5%
Omitted	4254	3%	151	4%
Multiple	65	0%	0	0%
Unknown	0	0%	0	0%
Math work is too easy				
Never or hardly eve	21759	14%	518	14%
Sometimes	77107	48%	1656	46%
Often	31769	20%	634	17%
Always or almost	24192	15%	648	18%
Omitted	4467	3%	167	5%
Multiple	44	0%	2	0%
Unknown	0	0%	0	0%
Like math				
Never or hardly eve	18905	12%	406	11%
Sometimes	38466	24%	794	22%
Often	33116	21%	680	19%
Always or almost	64083	40%	1567	43%

continued on next page...

...continued from previous page

	Charter		Traditional	
Omitted	4724	3%	177	5%
Multiple	44	0%	1	0%
Unknown	0	0%	0	0%

Table 11: Grade 4 Math Unadjusted NAEP Score

	Charter Schools						Public Schools					
	n	mean	sd	median	min	max	n	mean	sd	median	min	max
Overall	3625	231.93	28.19	232.18	136.95	310.45	159338	238.34	27.71	239.71	117.69	334.07
Alaska	102	247.41	24.64	248.79	183.80	296.82	2496	238.50	28.87	240.91	133.70	314.90
Arkansas	226	229.38	29.27	230.22	145.74	294.08	2836	228.66	29.47	229.71	142.09	324.86
Canal zone	170	223.41	26.86	221.21	152.40	290.14	7241	227.05	30.42	227.52	127.08	330.95
Connecticut	132	250.45	28.67	256.63	148.19	308.90	2550	242.34	28.39	244.89	130.83	317.99
D.C.	512	217.48	26.97	215.26	146.95	306.62	1282	219.99	32.15	219.20	133.26	317.28
Florida	195	240.20	26.34	238.76	168.45	294.30	2586	239.42	24.64	239.68	159.44	316.57
Idaho	166	241.09	24.28	244.36	175.09	293.17	4531	239.76	24.97	240.17	147.51	324.61
Illinois	164	241.86	25.56	244.89	183.18	310.45	3865	232.69	28.18	232.54	140.58	316.17
Iowa	57	230.95	26.26	234.88	146.85	273.46	2695	236.28	31.24	239.72	127.32	314.05
Kansas	60	255.08	20.79	255.56	183.42	300.83	3024	240.97	24.84	242.25	139.84	314.77
Kentucky	78	229.74	27.71	231.49	147.64	295.55	4059	232.94	29.83	233.75	117.69	318.75
Michigan	65	230.80	16.68	230.98	197.30	279.19	2830	229.16	24.90	229.32	144.26	308.71
Mississippi	141	219.12	26.52	217.25	164.93	293.18	3249	239.22	27.40	238.06	148.45	318.09
Missouri	52	229.29	22.62	226.30	179.65	284.75	3615	248.36	25.63	249.08	145.76	317.96
Montana	215	218.06	29.43	216.04	136.95	288.98	3204	229.38	31.66	231.15	128.36	318.70
Nebraska	109	247.68	26.51	248.08	188.30	302.32	3204	248.59	27.03	251.09	150.58	323.00
New York	72	231.01	31.20	229.16	162.07	303.29	2977	234.90	25.81	236.10	137.78	303.75
North Dakota	74	224.15	23.57	223.48	179.04	284.71	2780	247.30	26.13	249.06	155.28	325.63
Oklahoma	60	229.73	21.88	230.70	189.84	269.90	3997	239.50	26.83	240.45	145.65	315.10
Oregon	96	249.72	19.40	250.68	206.39	290.68	4320	243.07	26.92	243.12	130.47	323.33
Rhode Island	124	218.15	22.02	217.22	159.99	269.17	3324	237.62	29.58	239.43	126.61	319.60
Utah	79	230.94	23.07	227.57	185.38	297.38	2404	239.08	28.41	241.89	136.10	307.88
Washington	91	231.23	20.75	229.44	166.83	278.23	6193	238.97	24.48	239.19	139.25	317.06
West Virginia	197	247.20	25.79	250.06	164.01	305.53	3145	239.22	28.47	241.50	130.38	319.46
DoDEA/DDESS	136	229.43	26.14	226.78	178.41	289.56	3694	237.95	29.14	239.24	134.42	320.51

Table 12: Grade 4 Reading Descriptive Statistics

	Traditional		Charter	
Race/ethnicity from school records (raw data)				
White	96992	58%	1343	34%
Black	29127	17%	1636	42%
Hispanic	28133	17%	705	18%
Asian Amer/Pacif Is	8114	5%	162	4%
Amer Ind/Alaska Nat	3898	2%	49	1%
Other	2333	1%	41	1%
Unknown	0	0%	0	0%
Natl School Lunch Prog eligibility (3 categories)				
Eligible	82354	49%	2223	56%
Not eligible	85304	51%	1528	39%
Info not available	939	1%	185	5%
Unknown	0	0%	0	0%
Student has Individualized Education Plan				
Yes, IEP	16579	10%	307	8%
Yes, 504 plan	1385	1%	29	1%
Yes, 504 in process	0	0%	0	0%
Not IEP	150596	89%	3600	91%
Omitted	0	0%	0	0%
Unknown	37	0%	0	0%
Student classified Eng Lang Learner (3 categories)				
Yes	12095	7%	285	7%
No	153110	91%	3569	91%
Formerly ELL	3357	2%	82	2%
Omitted	0	0%	0	0%
Unknown	35	0%	0	0%
Gender				
Male	85214	51%	1960	50%
Female	83383	49%	1976	50%
Unknown	0	0%	0	0%
Student classified as having a disability (504)				
Student with disabi	17964	11%	336	9%
Not student with di	150596	89%	3600	91%
Omitted	37	0%	0	0%
Unknown	0	0%	0	0%
Student classified SD or ELL				
Student with disabi	16722	10%	314	8%
English language le	10853	6%	263	7%
Both SD and ELL	1242	1%	22	1%
Neither SD nor ELL	139727	83%	3337	85%
Unknown	53	0%	0	0%
Newspaper in home				
Yes	47839	28%	1141	29%
No	59247	35%	1327	34%
I Don't Know	58294	35%	1343	34%
Omitted	3205	2%	125	3%
Multiple	12	0%	0	0%

continued on next page...

...continued from previous page

	Charter		Traditional	
Unknown	0	0%	0	0%
Magazines in home				
Yes	95695	57%	2225	57%
No	40167	24%	911	23%
I Don't Know	29309	17%	667	17%
Omitted	3404	2%	133	3%
Multiple	22	0%	0	0%
Unknown	0	0%	0	0%
Books in home				
0-10 books	19634	12%	434	11%
11-25 books	35306	21%	901	23%
26-100 books	56725	34%	1217	31%
More than 100 books	53526	32%	1258	32%
Omitted	3359	2%	126	3%
Multiple	47	0%	0	0%
Unknown	0	0%	0	0%
Computer in home				
Yes	144162	86%	3362	85%
No	20419	12%	428	11%
Omitted	3999	2%	146	4%
Multiple	17	0%	0	0%
Unknown	0	0%	0	0%
Encyclopedia in home				
Yes	85818	51%	2035	52%
No	25320	15%	544	14%
I Don't Know	54087	32%	1229	31%
Omitted	3350	2%	128	3%
Multiple	22	0%	0	0%
Unknown	0	0%	0	0%
Pages read in school and for homework				
5 or fewer	34944	21%	912	23%
6-10	30880	18%	746	19%
11-15	23139	14%	449	11%
16-20	23805	14%	529	13%
More than 20	52313	31%	1167	30%
Omitted	3450	2%	129	3%
Multiple	66	0%	4	0%
Unknown	0	0%	0	0%
Talk about studies at home				
Never or hardly eve	29602	18%	648	16%
Every few weeks	22498	13%	475	12%
About once a week	19884	12%	429	11%
2-3 times a week	33893	20%	687	17%
Every day	59212	35%	1566	40%
Omitted	3465	2%	129	3%
Multiple	43	0%	2	0%
Unknown	0	0%	0	0%
Days absent from school last month				

continued on next page...

...continued from previous page

	Charter		Traditional	
None	84418	50%	1838	47%
1-2 days	49650	29%	1177	30%
3-4 days	19327	11%	474	12%
5-10 days	7608	5%	208	5%
More than 10 days	4136	2%	109	3%
Omitted	3396	2%	128	3%
Multiple	62	0%	2	0%
Unknown	0	0%	0	0%
Language other than English spoken in home				
Never	90390	54%	1798	46%
Once in a while	36078	21%	901	23%
Half the time	12161	7%	362	9%
All or most of time	26507	16%	743	19%
Omitted	3412	2%	129	3%
Multiple	49	0%	3	0%
Unknown	0	0%	0	0%
Learn a lot when reading books				
Never or hardly eve	8448	5%	185	5%
Sometimes	59897	36%	1331	34%
Often	50052	30%	1076	27%
Always or almost	46493	28%	1201	31%
Omitted	3687	2%	142	4%
Multiple	20	0%	1	0%
Unknown	0	0%	0	0%
Reading is a favorite subject				
Never or hardly eve	25581	15%	611	16%
Sometimes	60476	36%	1409	36%
Often	36703	22%	783	20%
Always or almost	41959	25%	987	25%
Omitted	3855	2%	146	4%
Multiple	23	0%	0	0%
Unknown	0	0%	0	0%
Do reading at after-school or tutoring program				
Yes	60364	36%	1718	44%
No	102803	61%	2029	52%
Omitted	5387	3%	189	5%
Multiple	43	0%	0	0%
Unknown	0	0%	0	0%
Go to book clubs, competitions, fairs for reading				
Yes	49006	29%	1255	32%
No	113968	68%	2491	63%
Omitted	5592	3%	189	5%
Multiple	31	0%	1	0%
Unknown	0	0%	0	0%
Read for fun on own				
Never or hardly eve	25028	15%	584	15%
Once or twice/month	24696	15%	569	14%
1-2 times a week	41186	24%	923	23%

continued on next page...

...continued from previous page

	Charter		Traditional	
Almost every day	72670	43%	1677	43%
Omitted	4984	3%	182	5%
Multiple	33	0%	1	0%
Unknown	0	0%	0	0%
Talk with friends about what you read				
Never or hardly eve	46333	27%	997	25%
Once or twice/month	34554	20%	739	19%
1-2 times a week	43383	26%	943	24%
Almost every day	40113	24%	1106	28%
Omitted	4180	2%	150	4%
Multiple	34	0%	1	0%
Unknown	0	0%	0	0%
Read a book you chose yourself				
Never or hardly eve	22712	13%	593	15%
Sometimes	40804	24%	1009	26%
Often	42467	25%	932	24%
Always or almost	56171	33%	1196	30%
Omitted	6413	4%	205	5%
Multiple	30	0%	1	0%
Unknown	0	0%	0	0%

Table 13: Grade 4 Reading Unadjusted NAEP Score

	Charter Schools						Public Schools					
	n	mean	sd	median	min	max	n	mean	sd	median	min	max
Overall	3936	213.26	32.94	215.22	92.27	302.97	168597	218.76	33.66	221.79	14.71	330.96
Alaska	117	226.06	33.86	233.08	101.93	278.72	2666	214.43	36.60	219.70	78.50	302.46
Arkansas	242	213.15	34.22	216.30	105.63	294.52	2943	208.58	38.27	212.92	14.71	319.32
Canal zone	183	196.96	35.30	194.82	92.27	286.62	7822	204.44	35.95	206.11	51.07	318.34
Connecticut	143	229.69	33.61	235.33	124.55	291.85	2777	225.49	34.52	230.29	85.98	316.87
D.C.	541	198.75	31.53	199.60	102.07	280.99	1307	204.14	37.69	203.42	71.14	330.96
Florida	216	222.89	27.69	223.23	152.52	293.06	2649	225.96	27.61	227.81	113.17	309.18
Idaho	185	227.25	24.58	228.11	141.86	288.14	4771	224.10	29.65	226.17	97.23	312.59
Illinois	179	220.16	32.54	220.82	94.63	287.95	4056	215.07	33.27	216.54	89.20	321.26
Iowa	78	215.14	36.63	222.41	96.15	289.85	2914	210.77	38.37	214.86	50.22	309.30
Kansas	72	242.29	27.73	247.02	157.89	290.30	3169	220.49	31.59	224.11	73.45	294.68
Kentucky	81	204.56	33.03	206.21	134.27	276.43	4314	213.09	36.59	216.16	82.70	322.40
Michigan	73	200.08	24.97	197.83	158.07	269.88	3079	207.87	31.56	209.30	75.54	298.16
Mississippi	152	207.05	28.29	208.95	142.04	274.30	3297	220.71	32.60	220.73	109.43	322.77
Missouri	54	225.77	22.09	225.23	169.00	265.05	3894	229.30	30.48	231.56	95.48	309.48
Montana	217	201.20	31.78	202.59	123.87	289.31	3487	212.74	34.23	215.29	75.67	302.63
Nebraska	109	221.06	36.91	223.34	101.89	285.11	3506	222.79	34.75	227.05	58.17	316.17
New York	75	215.17	29.94	216.60	134.65	278.63	3155	210.12	35.62	213.99	18.03	295.53
North Dakota	80	213.75	21.85	212.65	153.51	263.03	2805	230.06	28.90	231.41	118.16	321.05
Oklahoma	69	220.92	30.41	218.55	157.69	300.34	4162	221.95	32.22	224.25	99.34	306.82
Oregon	104	226.53	26.97	227.31	148.58	287.27	4720	219.46	34.73	223.00	40.55	309.55
Rhode Island	138	196.41	31.14	196.45	112.19	273.31	3464	218.63	32.97	220.65	101.22	305.91
South Dakota	56	225.22	31.87	229.31	148.54	285.83	3027	217.79	35.12	221.68	44.79	306.49
Tennessee	59	218.66	33.26	221.67	130.12	302.97	3846	216.38	36.39	218.94	45.75	316.07
Utah	85	217.99	31.30	218.14	135.33	283.96	2566	222.83	34.75	226.45	49.36	310.49
Washington	100	217.52	24.77	220.63	151.32	277.02	5854	216.69	30.85	216.91	91.15	318.03
West Virginia	203	228.59	27.26	231.71	137.46	287.80	3290	218.54	32.67	222.79	46.89	303.80
DoDEA/DDESS	153	200.04	32.60	198.92	123.03	273.93	3935	214.63	35.37	218.79	92.04	301.72

Table 14: Grade 8 Math Descriptive Statistics

	Traditional		Charter	
Race/ethnicity from school records (raw data)				
White	89701	59%	1114	27%
Black	26613	18%	1711	41%
Hispanic	23669	16%	974	24%
Asian Amer/Pacif Is	7318	5%	230	6%
Amer Ind/Alaska Nat	3250	2%	55	1%
Other	1497	1%	46	1%
Unknown	0	0%	0	0%
Natl School Lunch Prog eligibility (3 categories)				
Eligible	67525	44%	2358	57%
Not eligible	83452	55%	1553	38%
Info not available	1071	1%	219	5%
Unknown	0	0%	0	0%
Student has Individualized Education Plan				
Yes, IEP	14792	10%	377	9%
Yes, 504 plan	1308	1%	38	1%
Yes, 504 in process	0	0%	0	0%
Not IEP	135935	89%	3715	90%
Unknown	13	0%	0	0%
Student classified Eng Lang Learner (3 categories)				
Yes	6615	4%	276	7%
No	142006	93%	3712	90%
Formerly ELL	3404	2%	140	3%
Unknown	23	0%	2	0%
Gender				
Male	76976	51%	1996	48%
Female	75072	49%	2134	52%
Unknown	0	0%	0	0%
Student classified as having a disability (504)				
Student with disabi	16100	11%	415	10%
Not student with di	135935	89%	3715	90%
Omitted	13	0%	0	0%
Unknown	0	0%	0	0%
Student classified SD or ELL				
Student with disabi	15250	10%	389	9%
English language le	5765	4%	250	6%
Both SD and ELL	850	1%	26	1%
Neither SD nor ELL	130158	86%	3464	84%
Unknown	25	0%	1	0%
Newspaper in home				
Yes	55041	36%	1501	36%
No	62855	41%	1740	42%
I Don't Know	31056	20%	862	21%
Omitted	3068	2%	27	1%
Multiple	28	0%	0	0%
Unknown	0	0%	0	0%
Magazines in home				

continued on next page...

...continued from previous page

	Charter		Traditional	
Yes	92419	61%	2444	59%
No	40632	27%	1198	29%
I Don't Know	15801	10%	456	11%
Omitted	3172	2%	32	1%
Multiple	24	0%	0	0%
Unknown	0	0%	0	0%
Books in home				
0-10 books	21803	14%	578	14%
11-25 books	32216	21%	966	23%
26-100 books	51674	34%	1404	34%
More than 100 books	42985	28%	1148	28%
Omitted	3318	2%	34	1%
Multiple	52	0%	0	0%
Unknown	0	0%	0	0%
Computer in home				
Yes	133737	88%	3658	89%
No	13012	9%	386	9%
Omitted	5276	3%	86	2%
Multiple	23	0%	0	0%
Unknown	0	0%	0	0%
Encyclopedia in home				
Yes	106692	70%	3015	73%
No	22181	15%	571	14%
I Don't Know	19857	13%	509	12%
Omitted	3290	2%	35	1%
Multiple	28	0%	0	0%
Unknown	0	0%	0	0%
Pages read in school and for homework				
5 or fewer	44700	29%	1192	29%
6-10	32989	22%	928	22%
11-15	21654	14%	582	14%
16-20	17204	11%	512	12%
More than 20	31906	21%	865	21%
Omitted	3486	2%	46	1%
Multiple	109	0%	5	0%
Unknown	0	0%	0	0%
Talk about studies at home				
Never or hardly eve	35140	23%	825	20%
Every few weeks	28163	19%	776	19%
About once a week	26142	17%	698	17%
2-3 times a week	31254	21%	924	22%
Every day	27771	18%	863	21%
Omitted	3512	2%	44	1%
Multiple	66	0%	0	0%
Unknown	0	0%	0	0%
Days absent from school last month				
None	65078	43%	1692	41%
1-2 days	52510	35%	1393	34%

continued on next page...

...continued from previous page

	Charter		Traditional	
3-4 days	20084	13%	651	16%
5-10 days	7792	5%	257	6%
More than 10 days	3158	2%	101	2%
Omitted	3369	2%	35	1%
Multiple	57	0%	1	0%
Unknown	0	0%	0	0%
Mother's education level				
Did not finish h.s.	15175	10%	444	11%
Graduated h.s.	30320	20%	789	19%
Some ed after h.s.	25294	17%	752	18%
Graduated college	55231	36%	1396	34%
I Don't Know	22091	15%	701	17%
Omitted	3648	2%	44	1%
Multiple	289	0%	4	0%
Unknown	0	0%	0	0%
Father's education level				
Did not finish h.s.	15904	10%	457	11%
Graduated h.s.	30398	20%	730	18%
Some ed after h.s.	19878	13%	504	12%
Graduated college	46634	31%	1115	27%
I Don't Know	35054	23%	1267	31%
Omitted	3955	3%	52	1%
Multiple	225	0%	5	0%
Unknown	0	0%	0	0%
Language other than English spoken in home				
Never	86942	57%	1938	47%
Once in a while	28690	19%	860	21%
Half the time	11236	7%	428	10%
All or most of time	20361	13%	821	20%
Omitted	4766	3%	80	2%
Multiple	53	0%	3	0%
Unknown	0	0%	0	0%
Do math at after-school or tutoring program				
Yes	25981	17%	1025	25%
No	109053	72%	2712	66%
Omitted	16955	11%	393	10%
Multiple	59	0%	0	0%
Unknown	0	0%	0	0%
Math work is too easy				
Never or hardly eve	25733	17%	667	16%
Sometimes	79651	52%	2205	53%
Often	29995	20%	800	19%
Always/almost alway	11127	7%	339	8%
Omitted	5343	4%	111	3%
Multiple	199	0%	8	0%
Unknown	0	0%	0	0%
Math work is challenging				
Never or hardly eve	17626	12%	418	10%

continued on next page...

...continued from previous page

	Charter		Traditional	
Sometimes	64961	43%	1726	42%
Often	44818	29%	1298	31%
Always/almost alway	16629	11%	512	12%
Omitted	7824	5%	174	4%
Multiple	190	0%	2	0%
Unknown	0	0%	0	0%
Math work is engaging and interesting				
Never or hardly eve	34020	22%	756	18%
Sometimes	53378	35%	1415	34%
Often	38173	25%	1054	26%
Always or almost	19386	13%	732	18%
Omitted	7027	5%	171	4%
Multiple	64	0%	2	0%
Unknown	0	0%	0	0%
Math is fun				
Strongly disagree	17997	12%	472	11%
Disagree	49601	33%	1262	31%
Agree	62324	41%	1685	41%
Strongly agree	17723	12%	629	15%
Omitted	4319	3%	80	2%
Multiple	84	0%	2	0%
Unknown	0	0%	0	0%
Like math				
Strongly disagree	17227	11%	428	10%
Disagree	34661	23%	922	22%
Agree	69362	46%	1827	44%
Strongly agree	26051	17%	875	21%
Omitted	4628	3%	74	2%
Multiple	119	0%	4	0%
Unknown	0	0%	0	0%
Math is a favorite subject				
Strongly disagree	31790	21%	863	21%
Disagree	43981	29%	1133	27%
Agree	40525	27%	1020	25%
Strongly agree	30609	20%	1004	24%
Omitted	5108	3%	109	3%
Multiple	35	0%	1	0%
Unknown	0	0%	0	0%

Table 15: Grade 8 Math Unadjusted NAEP Score

	Charter Schools						Public Schools					
	n	mean	sd	median	min	max	n	mean	sd	median	min	max
Overall	4130	272.20	35.28	271.11	169.36	393.58	152048	280.77	34.88	281.50	126.93	400.47
Arkansas	105	273.39	34.78	265.78	169.36	343.91	2810	276.52	37.05	277.89	127.19	388.86
Canal zone	525	270.52	32.62	271.64	179.08	355.79	6606	266.15	36.73	265.25	146.91	384.02
Connecticut	125	294.37	39.89	294.63	184.51	393.58	2604	287.16	35.40	289.18	150.14	388.99
D.C.	825	256.83	30.99	254.59	173.87	346.00	870	251.82	39.61	250.26	142.88	384.80
Florida	201	294.07	34.66	294.89	204.19	374.46	2541	283.58	30.34	283.45	163.22	380.68
Idaho	272	283.12	29.62	282.81	192.00	370.99	4055	275.81	33.28	276.45	155.34	374.40
Illinois	90	262.43	29.52	262.06	193.38	330.81	3420	272.88	33.19	271.68	150.65	379.68
Iowa	167	278.44	33.09	281.78	183.00	354.42	2652	273.38	35.74	276.02	141.25	387.09
Kansas	90	302.88	29.24	303.84	198.35	377.42	2881	286.73	33.00	288.11	155.23	390.15
Kentucky	110	262.54	28.28	263.47	186.53	346.25	3981	276.59	34.83	276.44	145.80	377.44
Louisiana	77	283.53	31.97	290.89	206.16	341.36	2571	286.85	30.56	287.45	180.13	375.12
Michigan	90	266.23	32.25	264.51	193.70	359.20	2498	272.22	31.66	270.90	144.33	379.44
Mississippi	73	265.06	26.02	264.32	210.36	330.42	3132	280.86	37.37	279.87	165.91	380.51
Missouri	61	291.78	28.81	294.00	227.58	353.68	3513	294.51	36.13	295.84	160.48	390.57
Montana	160	250.50	32.19	248.72	182.79	361.40	3221	268.85	39.61	269.65	152.12	381.41
Nebraska	80	296.78	37.39	299.59	186.61	371.70	2818	293.76	33.17	295.30	176.56	388.60
Oregon	93	285.52	33.31	287.31	201.69	346.33	4347	281.90	36.25	282.12	138.57	399.87
Rhode Island	94	264.65	29.71	260.31	205.85	355.42	3400	278.90	34.16	278.95	159.93	383.14
Tennessee	112	266.91	31.22	268.06	185.07	329.18	3439	282.63	36.14	284.27	141.96	391.93
Washington	134	291.15	37.59	291.27	199.06	372.41	5650	283.48	33.17	283.10	147.24	392.43
West Virginia	118	294.04	32.96	298.47	221.08	366.48	2765	282.25	33.24	284.52	153.63	374.36
DoDEA/DDESS	231	252.79	31.72	251.29	178.28	376.65	3243	282.00	34.97	284.13	143.51	372.36

Table 16: Grade 8 Reading Descriptive Statistics

	Traditional		Charter	
Race/ethnicity from school records (raw data)				
White	89855	59%	1147	28%
Black	26163	17%	1631	40%
Hispanic	23219	15%	974	24%
Asian Amer/Pacif Is	7232	5%	241	6%
Amer Ind/Alaska Nat	3341	2%	50	1%
Other	1494	1%	45	1%
Unknown	0	0%	0	0%
Natl School Lunch Prog eligibility (3 categories)				
Eligible	66739	44%	2282	56%
Not eligible	83449	55%	1593	39%
Info not available	1116	1%	213	5%
Unknown	0	0%	0	0%
Student has Individualized Education Plan				
Yes, IEP	13779	9%	334	8%
Yes, 504 plan	1433	1%	38	1%
Yes, 504 in process	0	0%	0	0%
Not IEP	136080	90%	3714	91%
Unknown	12	0%	2	0%
Student classified Eng Lang Learner (3 categories)				
Yes	5609	4%	278	7%
No	142262	94%	3674	90%
Formerly ELL	3422	2%	132	3%
Unknown	11	0%	4	0%
Gender				
Male	76149	50%	1887	46%
Female	75155	50%	2201	54%
Unknown	0	0%	0	0%
Student classified as having a disability (504)				
Student with disabi	15212	10%	372	9%
Not student with di	136080	90%	3714	91%
Omitted	12	0%	2	0%
Unknown	0	0%	0	0%
Student classified SD or ELL				
Student with disabi	14453	10%	343	8%
English language le	4850	3%	249	6%
Both SD and ELL	759	1%	29	1%
Neither SD nor ELL	131226	87%	3463	85%
Unknown	16	0%	4	0%
Newspaper in home				
Yes	54092	36%	1448	35%
No	63146	42%	1765	43%
I Don't Know	31080	21%	842	21%
Omitted	2962	2%	33	1%
Multiple	24	0%	0	0%
Unknown	0	0%	0	0%
Magazines in home				

continued on next page...

...continued from previous page

	Charter		Traditional	
Yes	92551	61%	2405	59%
No	39842	26%	1194	29%
I Don't Know	15813	10%	455	11%
Omitted	3079	2%	34	1%
Multiple	19	0%	0	0%
Unknown	0	0%	0	0%
Books in home				
0-10 books	20713	14%	567	14%
11-25 books	31676	21%	960	23%
26-100 books	52567	35%	1392	34%
More than 100 books	43159	29%	1134	28%
Omitted	3145	2%	35	1%
Multiple	44	0%	0	0%
Unknown	0	0%	0	0%
Computer in home				
Yes	133345	88%	3639	89%
No	12521	8%	351	9%
Omitted	5411	4%	97	2%
Multiple	27	0%	1	0%
Unknown	0	0%	0	0%
Encyclopedia in home				
Yes	105951	70%	2979	73%
No	21837	14%	542	13%
I Don't Know	20295	13%	527	13%
Omitted	3196	2%	40	1%
Multiple	25	0%	0	0%
Unknown	0	0%	0	0%
Pages read in school and for homework				
5 or fewer	42569	28%	1103	27%
6-10	33717	22%	943	23%
11-15	22450	15%	609	15%
16-20	18014	12%	497	12%
More than 20	31097	21%	890	22%
Omitted	3365	2%	44	1%
Multiple	92	0%	2	0%
Unknown	0	0%	0	0%
Talk about studies at home				
Never or hardly eve	33589	22%	798	20%
Every few weeks	27196	18%	723	18%
About once a week	26128	17%	693	17%
2-3 times a week	32644	22%	922	23%
Every day	28300	19%	907	22%
Omitted	3401	2%	43	1%
Multiple	46	0%	2	0%
Unknown	0	0%	0	0%
Days absent from school last month				
None	64800	43%	1734	42%
1-2 days	52913	35%	1411	35%

continued on next page...

...continued from previous page

	Charter		Traditional	
3-4 days	19761	13%	580	14%
5-10 days	7545	5%	239	6%
More than 10 days	2959	2%	85	2%
Omitted	3273	2%	37	1%
Multiple	53	0%	2	0%
Unknown	0	0%	0	0%
Mother's education level				
Did not finish h.s.	14613	10%	404	10%
Graduated h.s.	29684	20%	738	18%
Some ed after h.s.	25710	17%	780	19%
Graduated college	55743	37%	1439	35%
I Don't Know	21753	14%	670	16%
Omitted	3557	2%	46	1%
Multiple	244	0%	11	0%
Unknown	0	0%	0	0%
Father's education level				
Did not finish h.s.	15332	10%	423	10%
Graduated h.s.	30398	20%	710	17%
Some ed after h.s.	20427	14%	545	13%
Graduated college	46954	31%	1153	28%
I Don't Know	34150	23%	1197	29%
Omitted	3868	3%	53	1%
Multiple	175	0%	7	0%
Unknown	0	0%	0	0%
Language other than English spoken in home				
Never	86786	57%	1894	46%
Once in a while	28564	19%	922	23%
Half the time	11228	7%	433	11%
All or most of time	20222	13%	765	19%
Omitted	4455	3%	72	2%
Multiple	49	0%	2	0%
Unknown	0	0%	0	0%
Reading is a favorite activity				
Strongly disagree	37705	25%	830	20%
Disagree	54254	36%	1467	36%
Agree	33790	22%	1103	27%
Strongly agree	19268	13%	579	14%
Omitted	6237	4%	108	3%
Multiple	50	0%	1	0%
Unknown	0	0%	0	0%
Read for fun on own				
Never or hardly eve	46454	31%	1032	25%
Once or twice/month	33556	22%	1021	25%
1-2 times a week	35520	23%	1055	26%
Almost every day	30931	20%	902	22%
Omitted	4769	3%	76	2%
Multiple	74	0%	2	0%
Unknown	0	0%	0	0%

continued on next page...

...continued from previous page

	Charter		Traditional	
Use school/public library for info for own use				
Never or hardly eve	76861	51%	2100	51%
Once/twice a month	44184	29%	1205	29%
Once or twice a wee	20071	13%	558	14%
Every day or almost	6138	4%	175	4%
Omitted	4029	3%	50	1%
Multiple	21	0%	0	0%
Unknown	0	0%	0	0%
Do Eng/lang arts at after-school or tutoring prog				
Yes	26604	18%	1099	27%
No	120344	80%	2920	71%
Omitted	4330	3%	67	2%
Multiple	26	0%	2	0%
Unknown	0	0%	0	0%
Go to book clubs, competitions, fairs for reading				
Yes	33064	22%	1147	28%
No	111689	74%	2818	69%
Omitted	6523	4%	123	3%
Multiple	28	0%	0	0%
Unknown	0	0%	0	0%

Table 17: Grade 8 Reading Unadjusted NAEP Score

	Charter Schools						Public Schools					
	n	mean	sd	median	min	max	n	mean	sd	median	min	max
Overall	4088	256.27	32.94	258.06	122.77	350.97	151304	261.65	31.88	264.69	73.88	395.38
Arkansas	96	256.04	41.63	256.14	122.77	339.36	2739	256.81	34.43	260.57	126.56	342.91
Canal zone	500	248.96	35.63	252.66	126.10	345.55	6684	247.48	35.33	250.24	96.30	358.39
Connecticut	148	276.04	28.27	283.61	163.67	341.03	2607	264.17	30.14	266.49	124.53	356.84
D.C.	792	245.50	29.44	246.59	148.27	330.85	826	241.77	37.62	241.67	124.79	340.51
Florida	195	273.34	30.08	275.57	190.41	340.18	2557	264.75	27.37	267.22	166.92	332.18
Idaho	281	268.82	26.71	271.59	192.71	331.07	3928	262.15	31.40	264.16	133.45	349.50
Illinois	86	247.23	31.37	250.03	159.75	307.27	3397	258.36	31.06	259.82	119.83	343.14
Iowa	172	256.51	32.18	257.96	173.44	331.34	2693	254.25	31.71	257.49	133.60	337.56
Kansas	86	284.32	25.90	289.36	194.44	341.67	2879	263.96	30.26	267.31	116.64	336.22
Kentucky	107	248.44	23.29	250.67	198.59	300.52	3996	259.71	32.04	262.56	144.33	349.01
Louisiana	77	253.69	20.75	256.41	207.58	294.13	2579	266.40	28.59	268.56	140.72	337.32
Michigan	89	250.34	27.67	250.06	170.25	333.93	2514	253.14	31.67	255.03	111.28	348.55
Mississippi	77	247.56	21.96	245.11	196.99	290.06	3095	262.49	32.50	263.18	115.34	352.86
Missouri	56	276.37	29.72	272.99	182.71	343.64	3557	269.19	31.68	271.05	132.08	350.95
Montana	162	244.84	30.73	244.12	154.60	319.40	3174	255.36	34.99	258.19	113.45	349.25
Nebraska	78	272.22	39.25	280.31	174.13	350.97	2803	269.01	29.01	272.52	117.90	350.65
Oregon	90	266.50	37.32	272.90	170.18	344.88	4374	257.97	34.14	261.08	98.30	357.58
Rhode Island	87	253.08	28.29	255.80	193.19	304.42	3267	262.96	31.65	265.72	112.27	344.09
Tennessee	118	255.62	30.19	254.58	191.20	320.67	3429	263.73	32.48	266.63	145.83	358.89
Washington	131	265.94	34.58	271.95	147.12	335.37	5602	257.67	31.76	260.22	102.43	341.04
West Virginia	118	275.83	27.11	277.67	189.22	326.77	2712	263.53	30.14	266.53	143.86	340.59
DoDEA/DDESS	216	245.73	31.56	246.19	159.34	337.02	3181	262.14	32.15	266.18	119.62	339.34

Appendix C

Covariate Missingness

Figure 16: Covariate Missingness for Grade 4 Math

Figure 17: Covariate Missingness for Grade 4 Reading

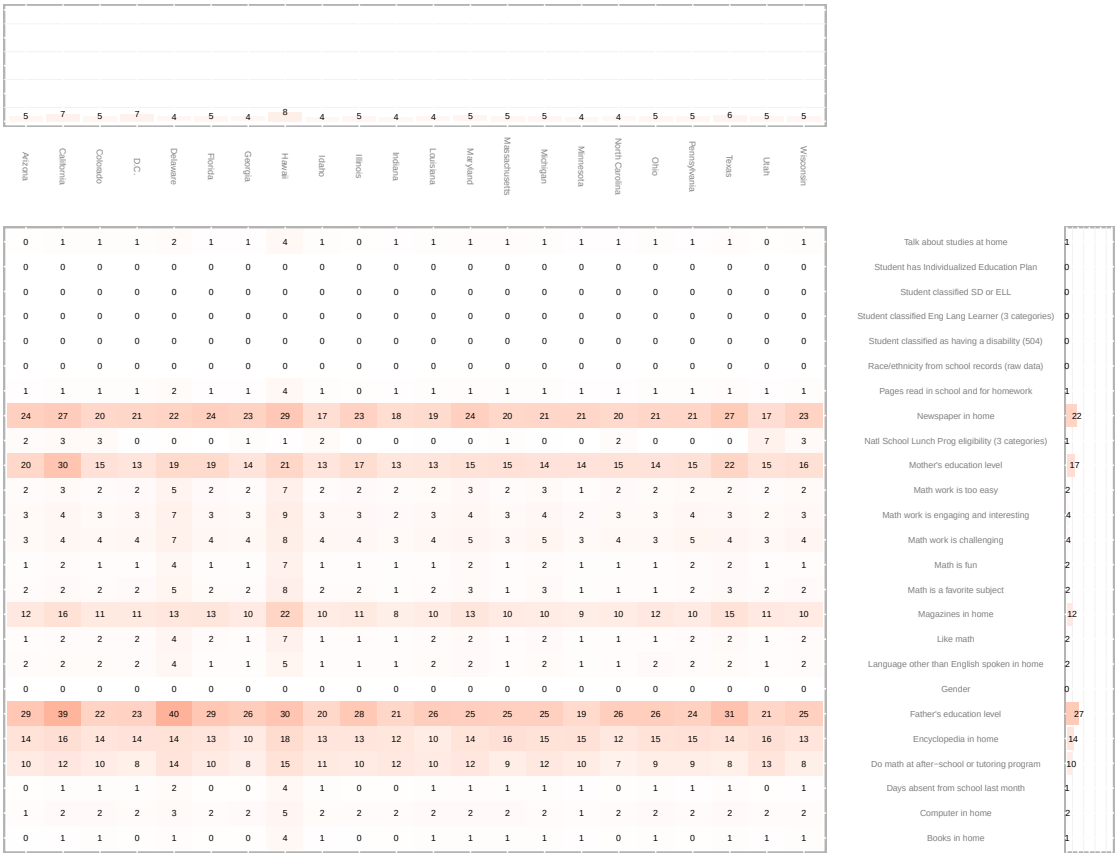


Figure 18: Covariate Missingness for Grade 8 Math

Figure 19: Covariate Missingness for Grade 8 Reading

Appendix D

Loess Regression Plots

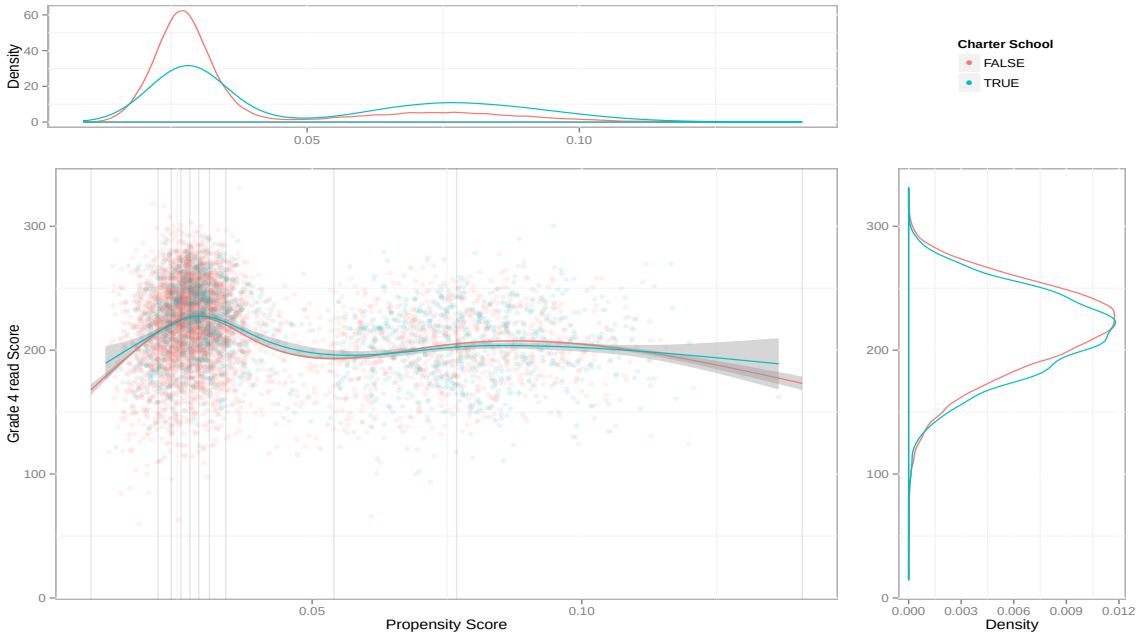


Figure 20: Loess Regression Assessment Plot: Grade 4 Reading

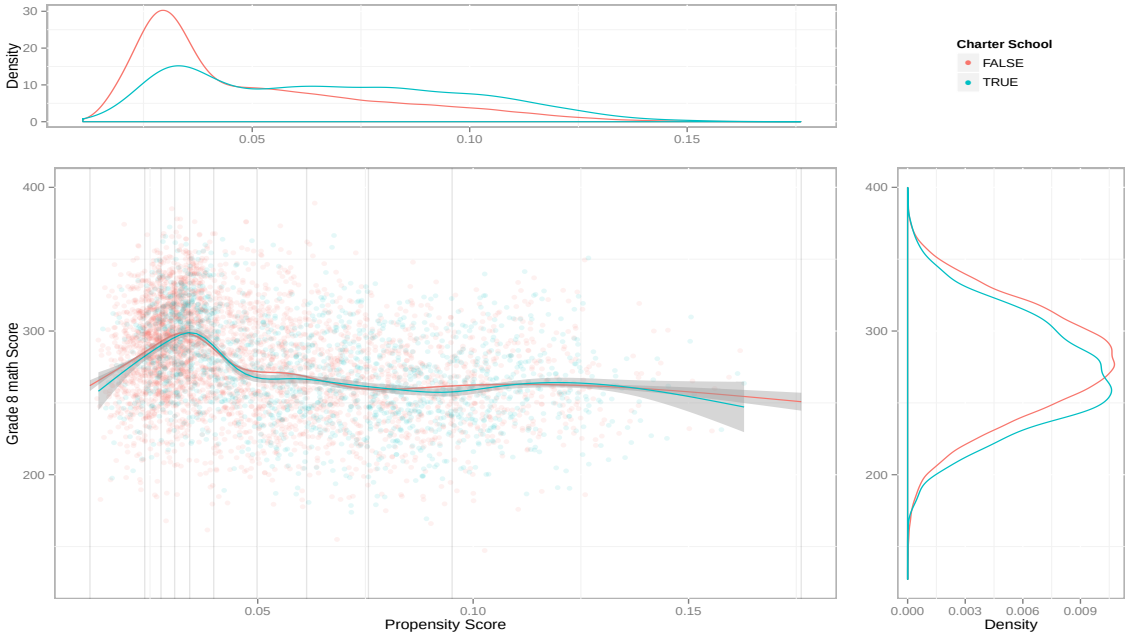


Figure 21: Loess Regression Assessment Plot: Grade 8 Math

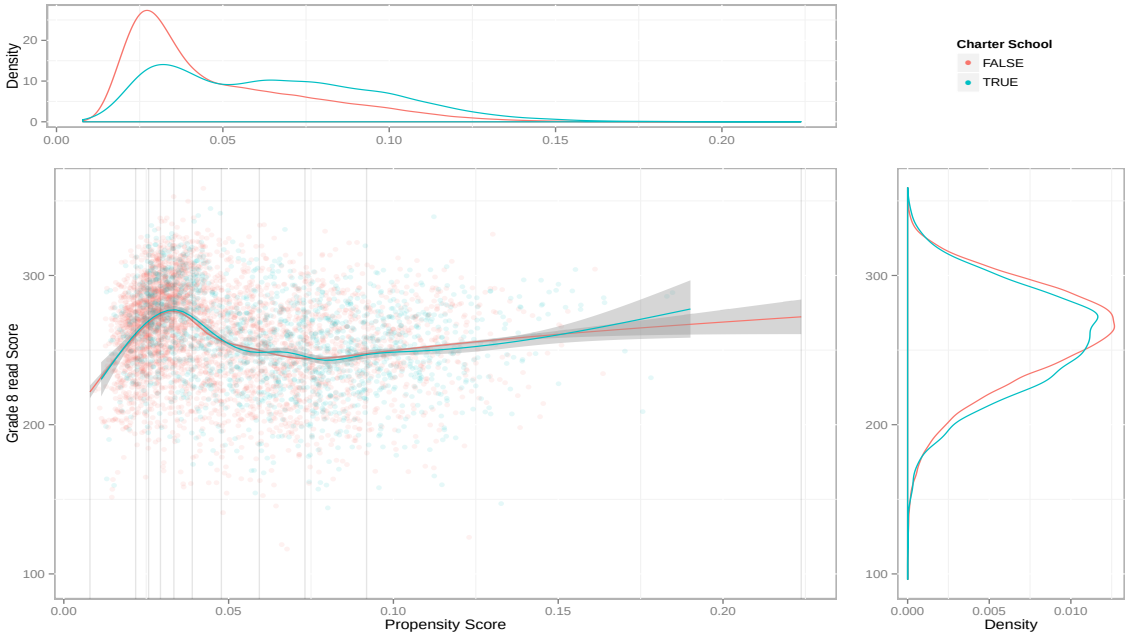


Figure 22: Loess Regression Assessment Plot: Grade 8 Reading

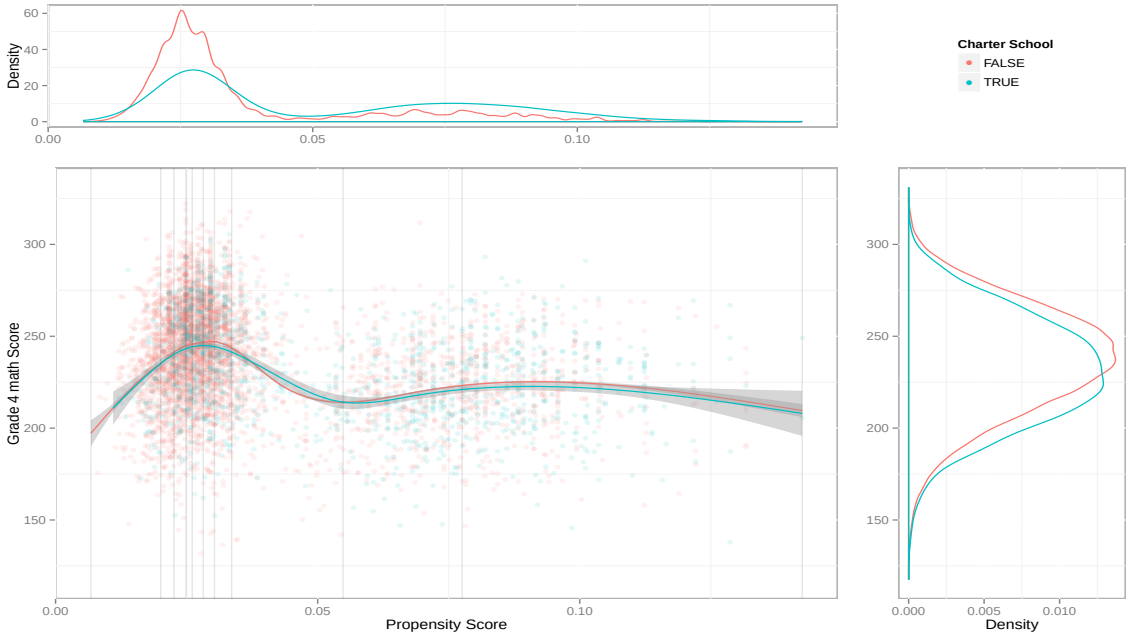


Figure 23: Loess Regression AIC Assessment Plot: Grade 4 Math

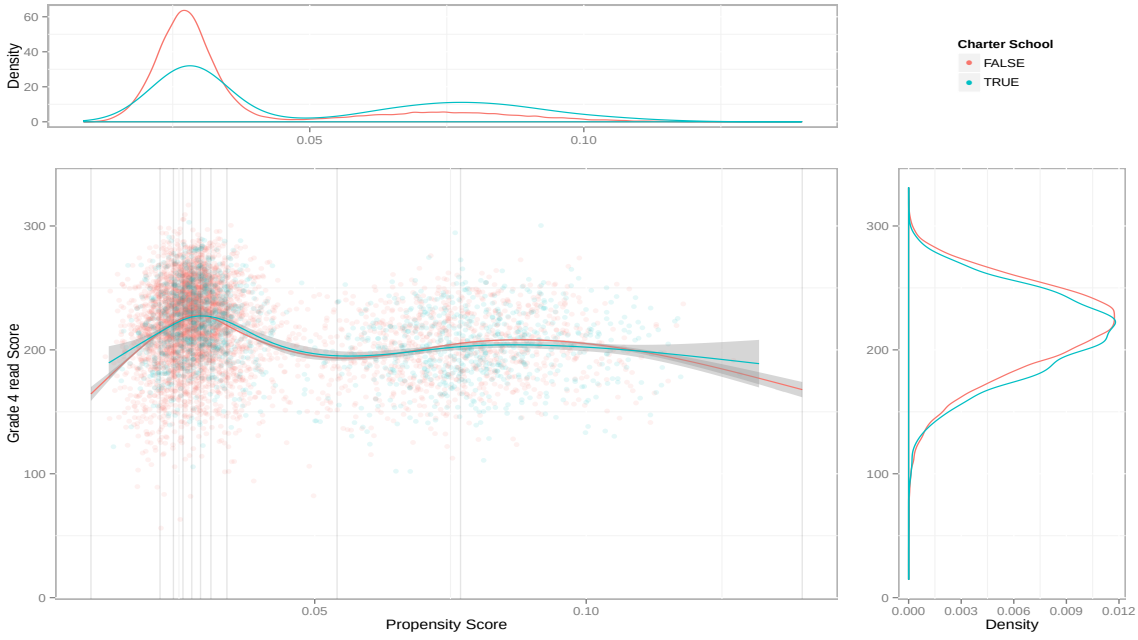


Figure 24: Loess Regression AIC Assessment Plot: Grade 4 Reading

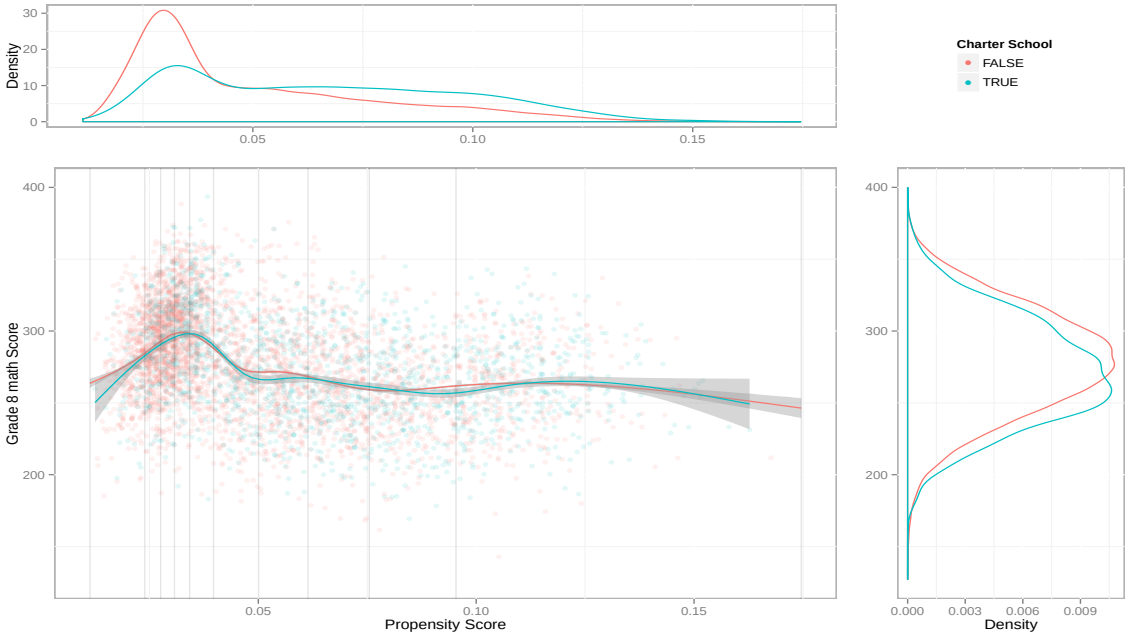


Figure 25: Loess Regression AIC Assessment Plot: Grade 8 Math

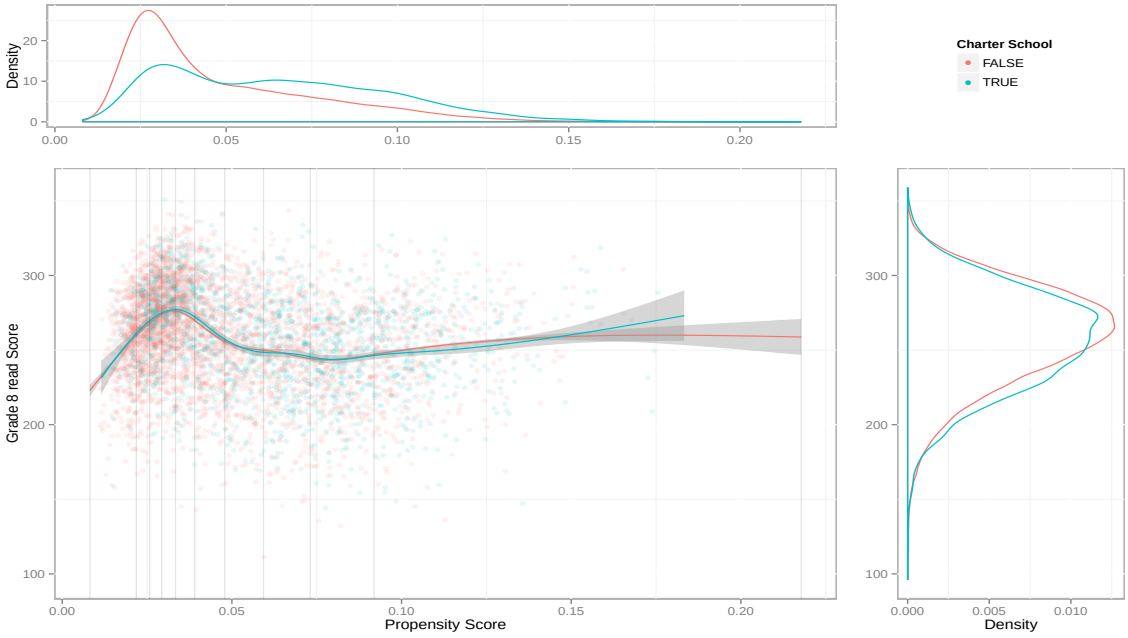


Figure 26: Loess Regression AIC Assessment Plot: Grade 8 Reading

Appendix E

Covariate Balance Plots

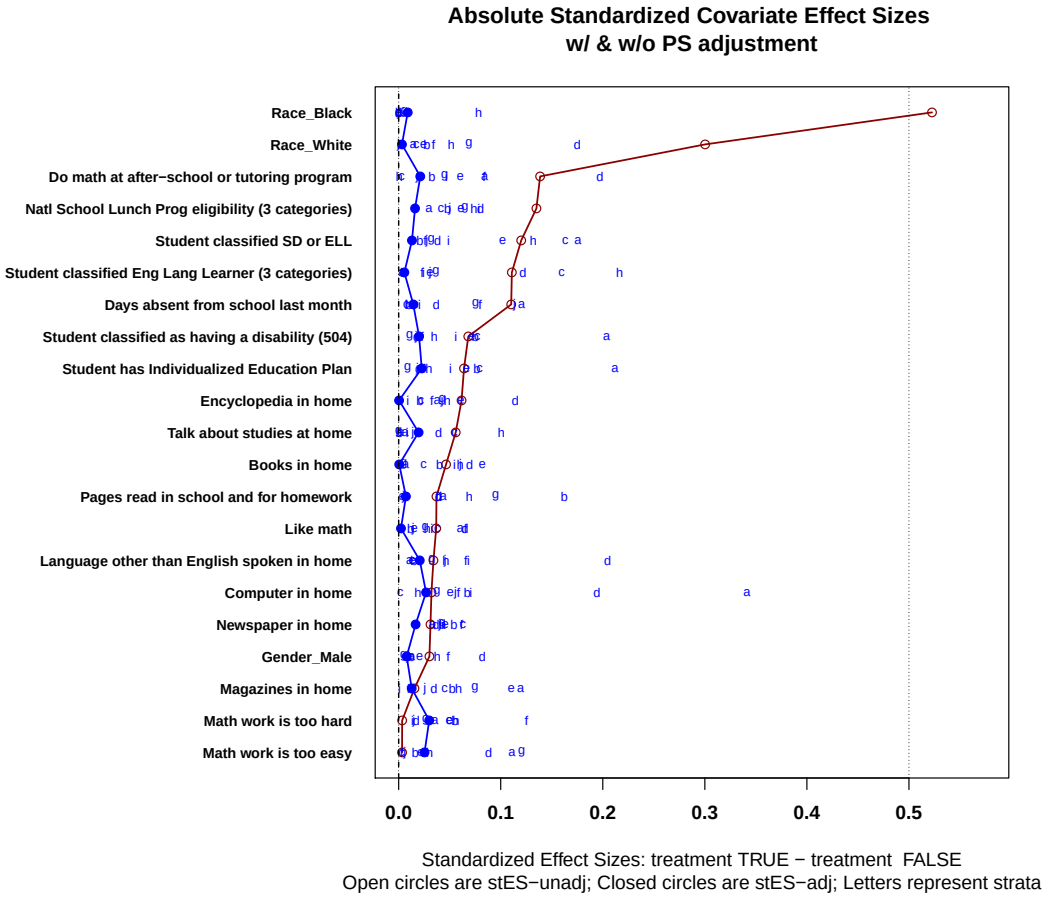


Figure 27: Covariate Balance Plot for Logistic Regression AIC Stratification: Grade 4 Math

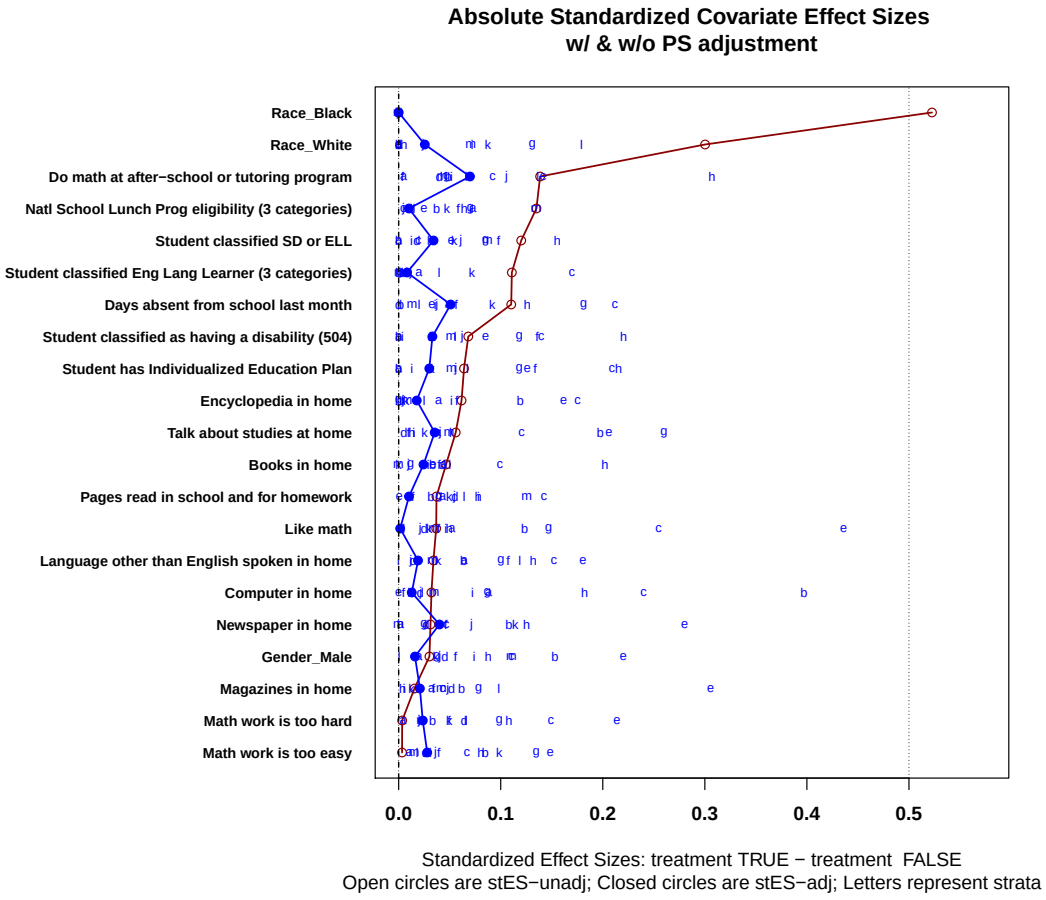


Figure 28: Covariate Balance Plot for Classification Tree Stratification: Grade 4 Math

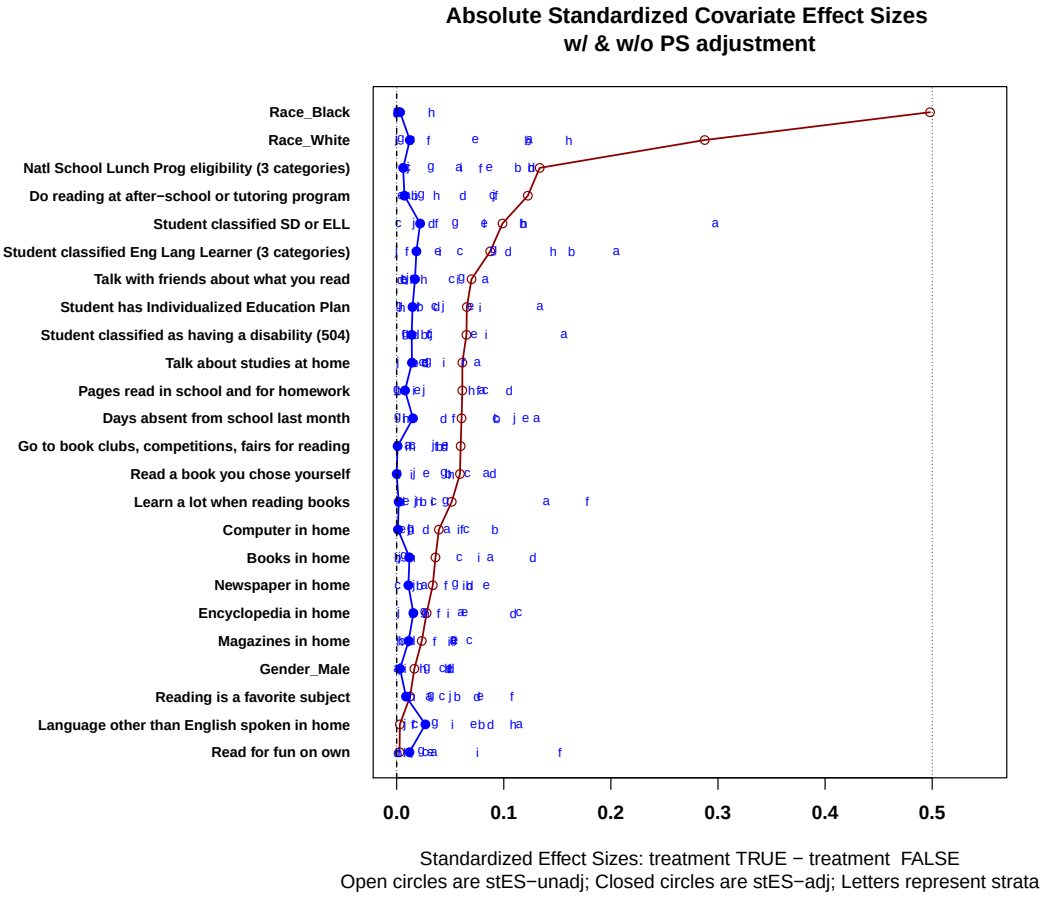


Figure 29: Covariate Balance Plot for Logistic Regression Stratification: Grade 4 Reading

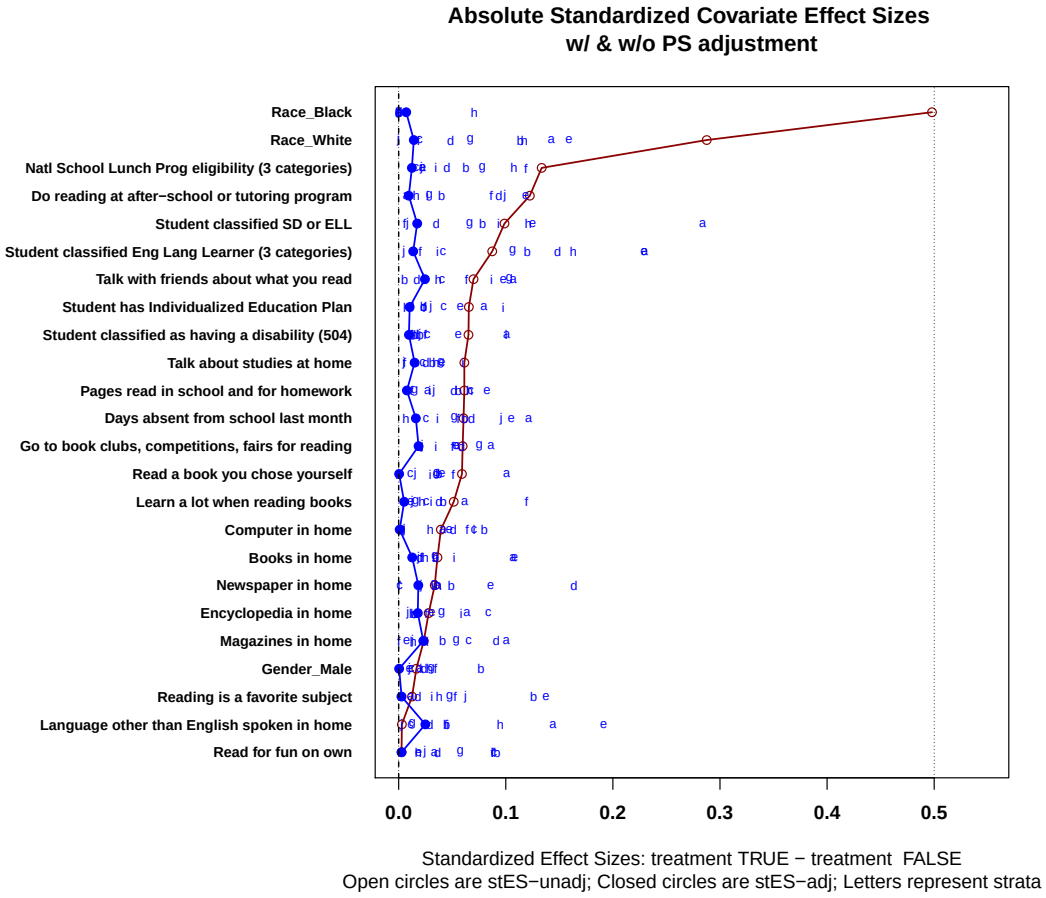


Figure 30: Covariate Balance Plot for Logistic Regression AIC Stratification: Grade 4 Reading

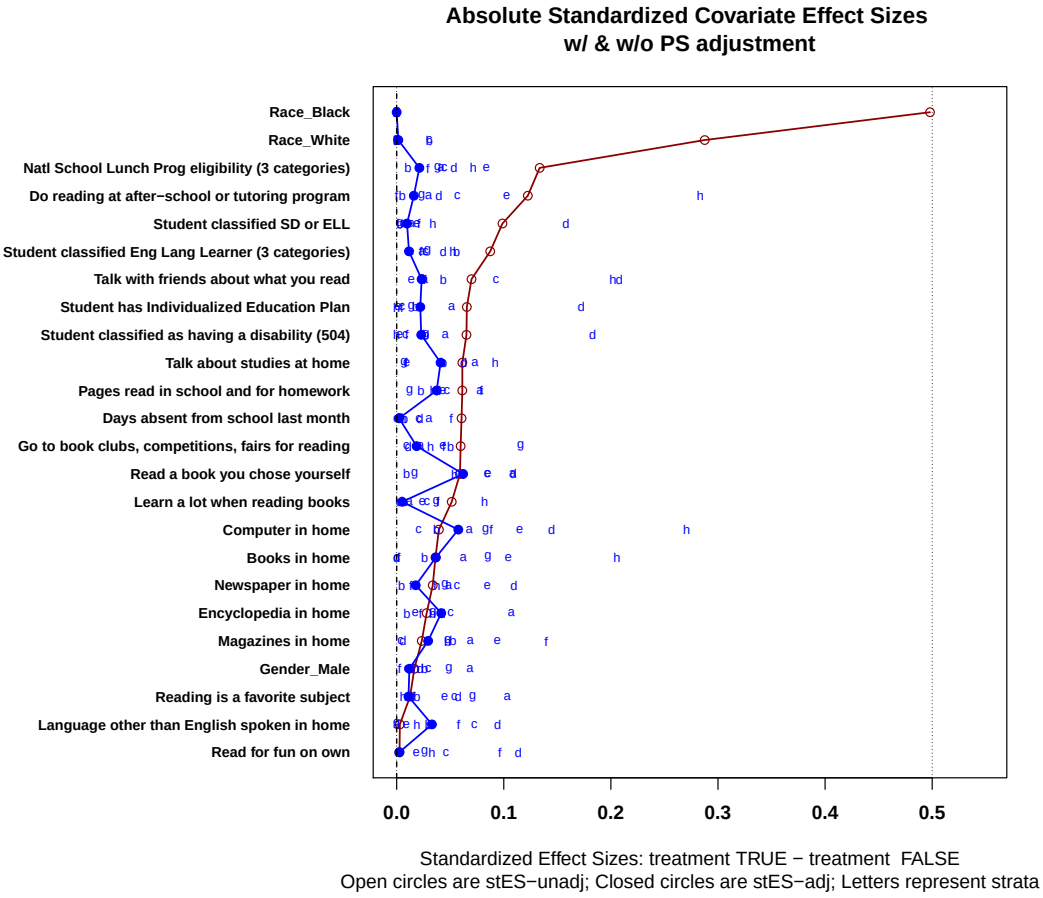


Figure 31: Covariate Balance Plot for Classification Tree Stratification: Grade 4 Reading

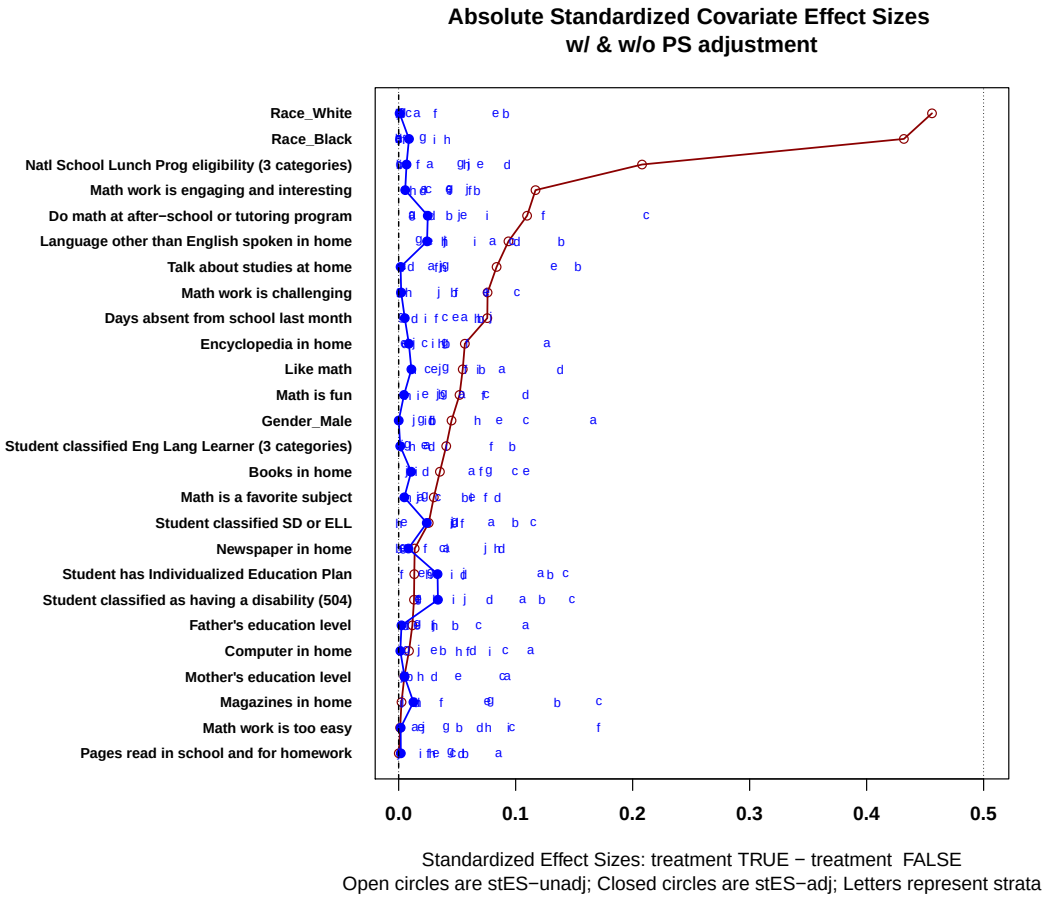


Figure 32: Covariate Balance Plot for Logistic Regression Stratification: Grade 8 Math

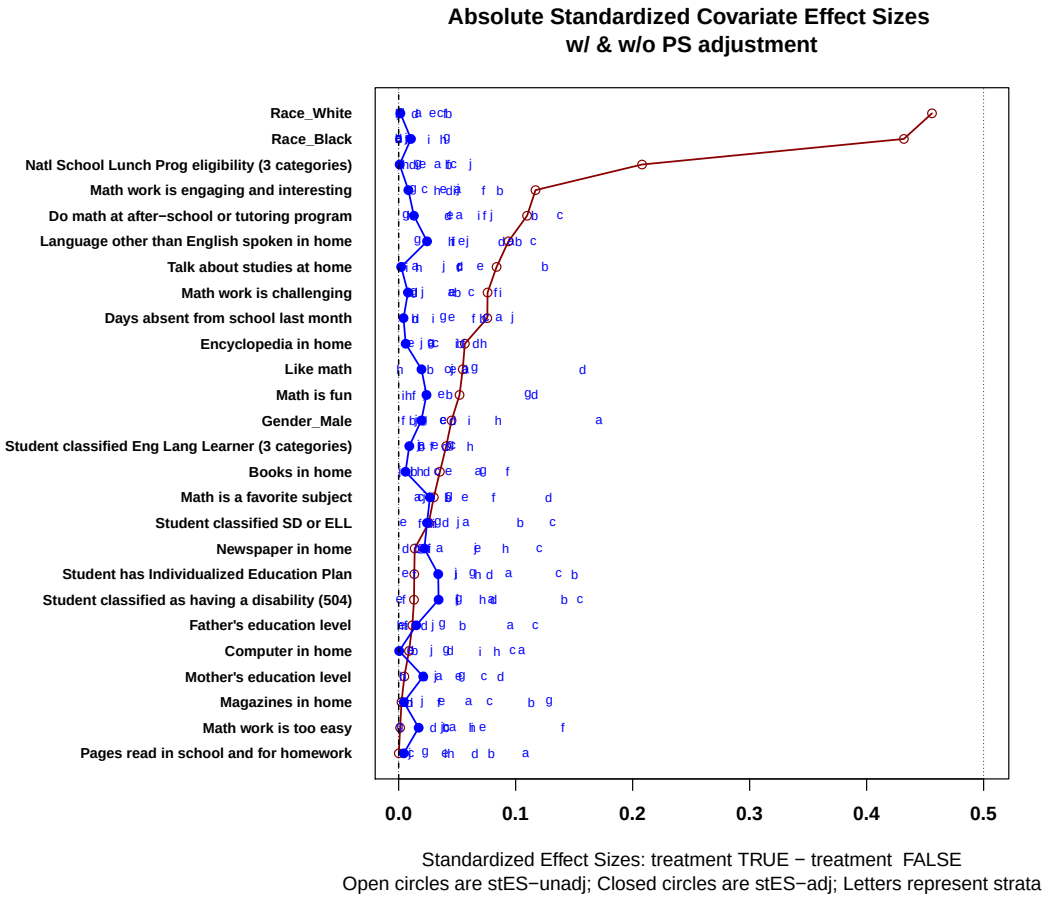


Figure 33: Covariate Balance Plot for Logistic Regression AIC Stratification: Grade 8 Math

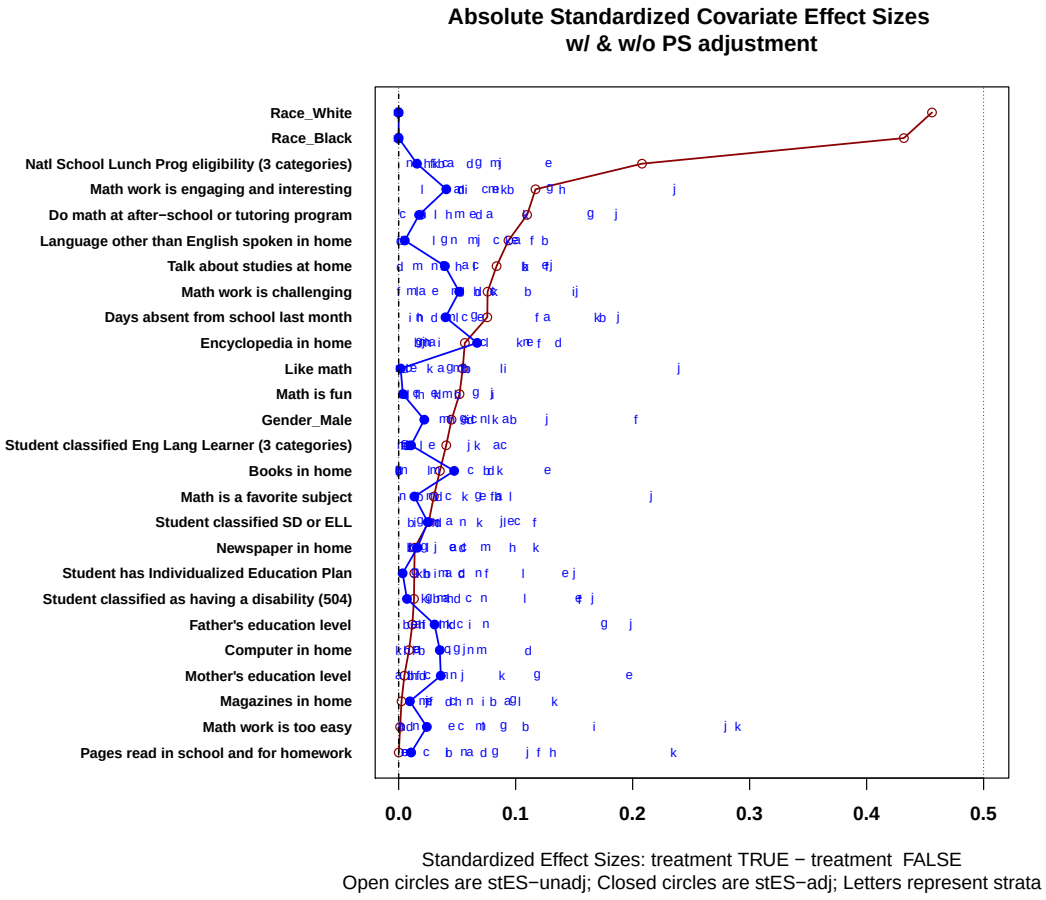


Figure 34: Covariate Balance Plot for Classification Tree Stratification: Grade 8 Math

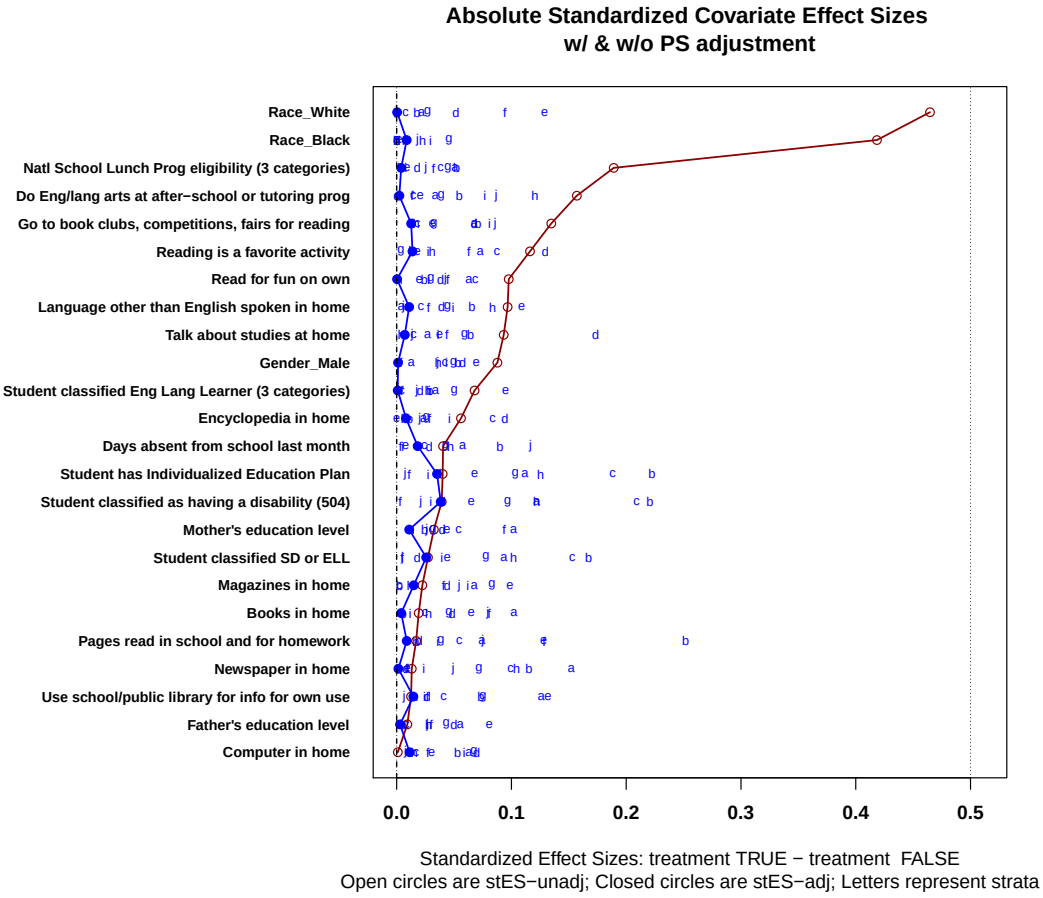


Figure 35: Covariate Balance Plot for Logistic Regression Stratification: Grade 8 Reading

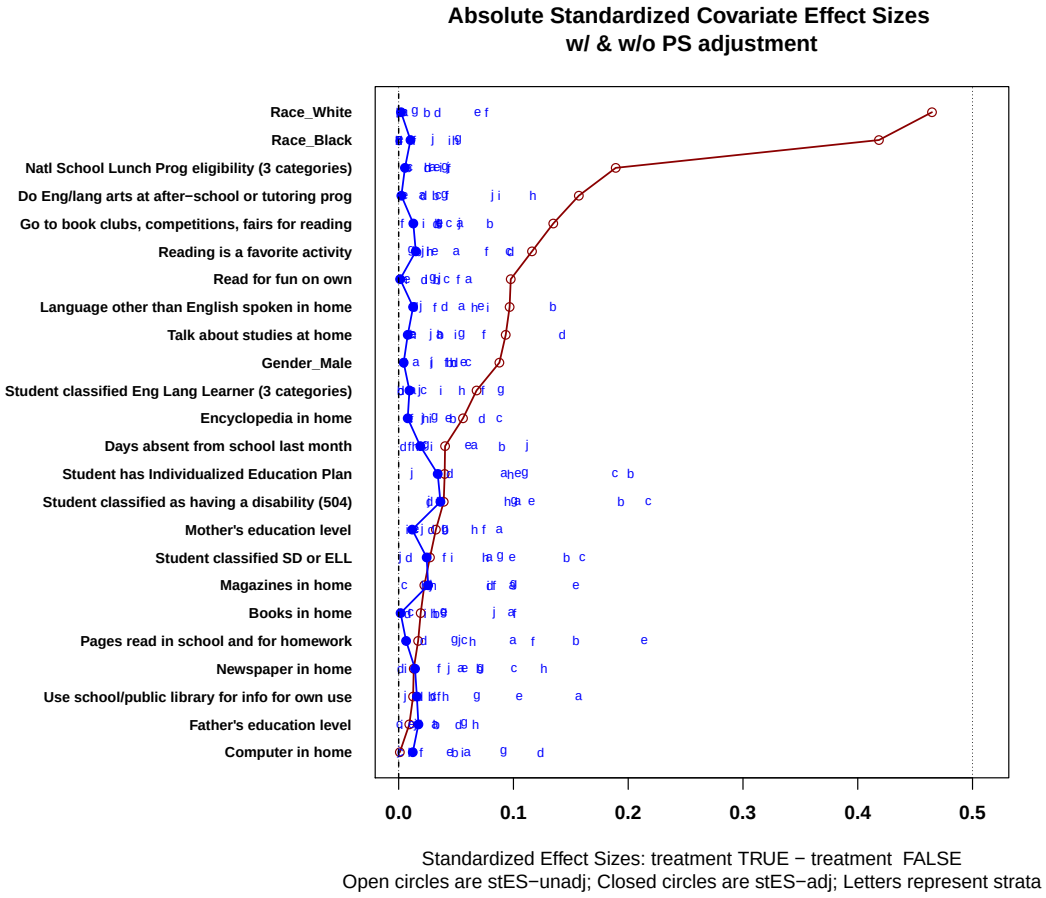


Figure 36: Covariate Balance Plot for Logistic Regression AIC Stratification: Grade 8 Reading

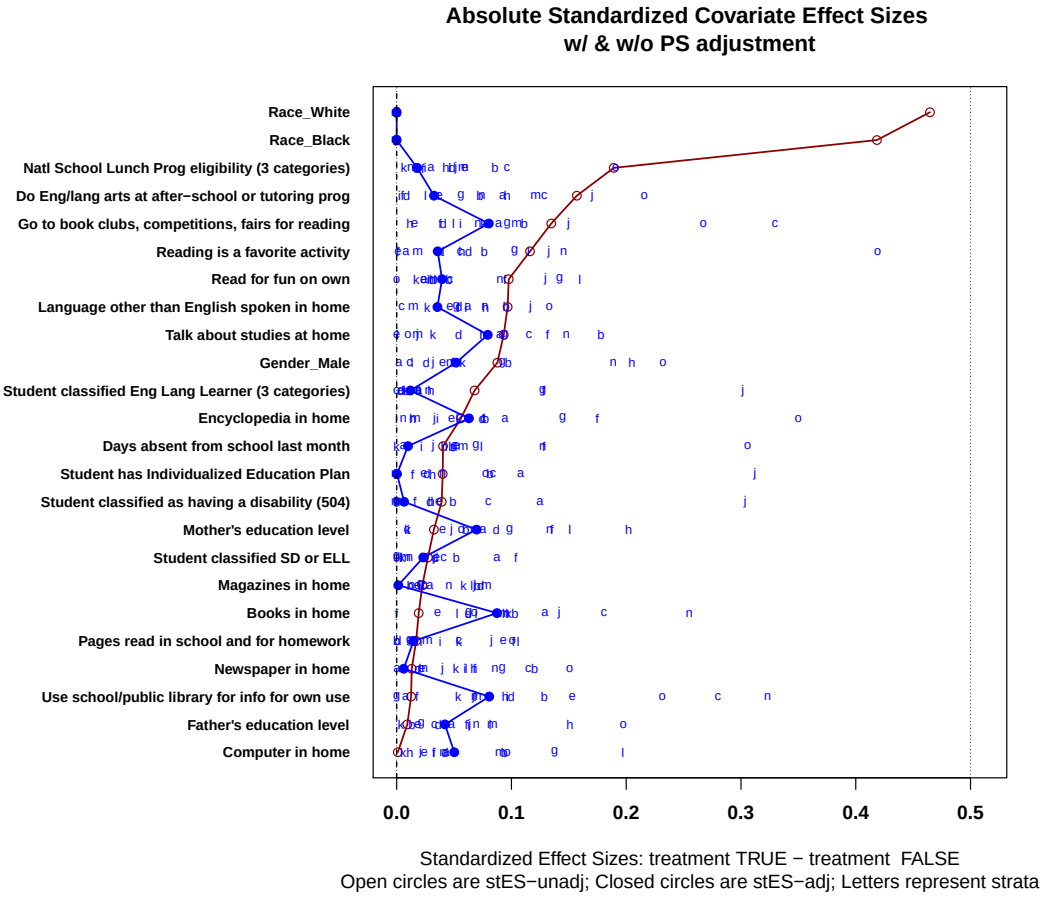


Figure 37: Covariate Balance Plot for Classification Tree Stratification: Grade 8 Reading

Appendix F

Classification Method Results

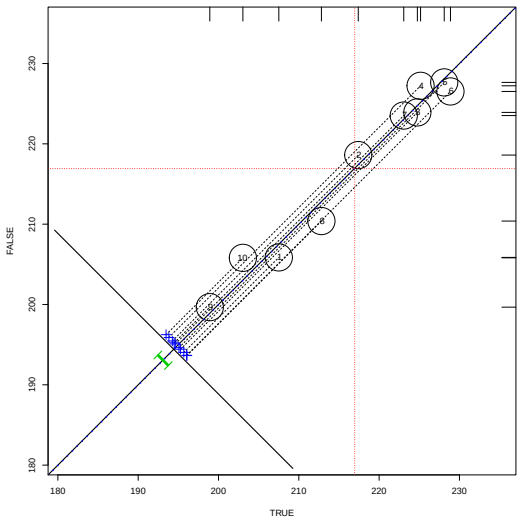


Figure 38: Propensity Score Assessment Plot for Logistic Regression Stratification: Grade 4 Reading

Table 18: Logistic Regression Stratification Results for Grade 4 read

	Public		Charter	
Strata	Mean	n	Mean	n
1	205.86	9440	207.51	201
2	218.60	9396	217.40	244
3	223.91	9439	224.76	202
4	227.23	9403	225.16	236
5	227.65	9371	228.10	270
6	226.52	9334	228.88	306
7	223.52	9317	223.07	323
8	210.39	9292	212.81	348
9	199.67	8990	198.92	650
10	205.81	8774	203.04	867

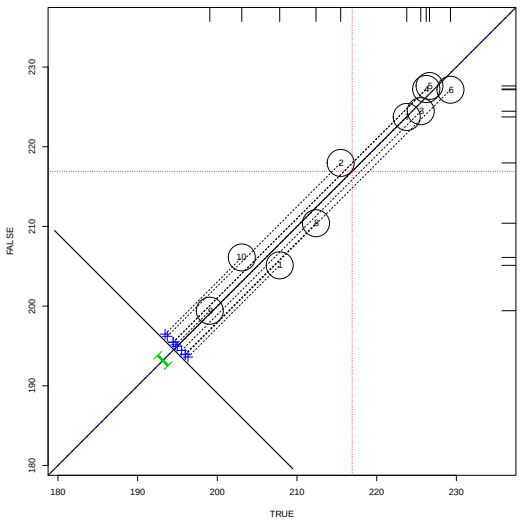


Figure 39: Propensity Score Assessment Plot for Logistic Regression AIC Stratification: Grade 4 Reading

Table 19: Logistic Regression AIC Stratification Results for Grade 4 read

	Public		Charter	
Strata	Mean	n	Mean	n
1	205.11	9434	207.83	218
2	217.96	9415	215.48	215
3	224.45	9406	225.53	233
4	227.24	9562	226.22	230
5	227.62	9241	226.63	263
6	227.15	9341	229.26	284
7	223.73	9296	223.77	354
8	210.40	9293	212.38	337
9	199.42	8996	199.06	648
10	206.10	8772	203.07	865

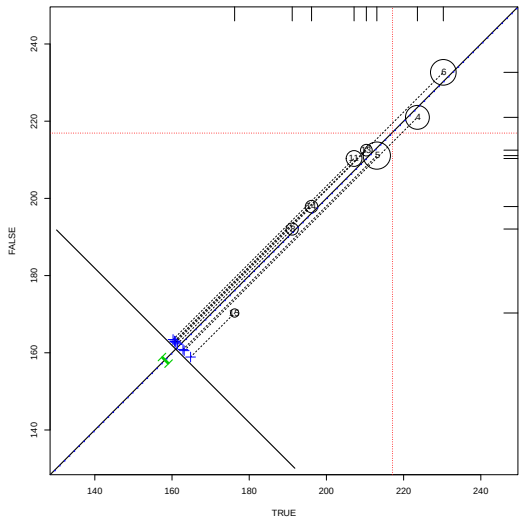


Figure 40: Propensity Score Assessment Plot for Classification Tree Stratification: Grade 4 Reading

Table 20: Classification Trees Stratification Results for Grade 4 read

	Public		Charter	
Strata	Mean	n	Mean	n
4	220.99	20677	223.57	464
5	211.08	28477	213.06	826
6	232.65	24503	230.28	785
8	192.06	3690	191.14	223
11	210.33	7001	207.16	547
13	212.50	3225	210.35	294
14	197.88	3735	196.15	429
15	170.31	1448	176.22	79

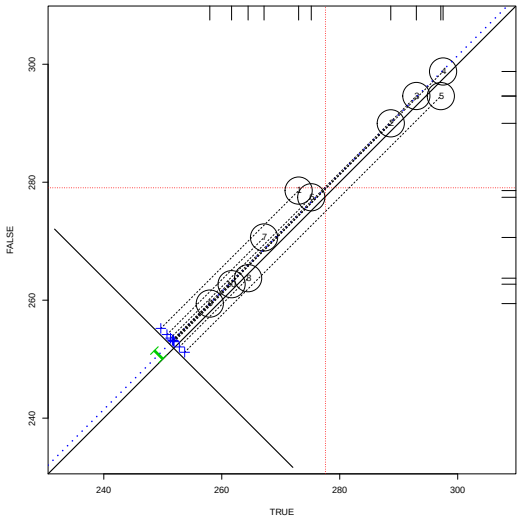


Figure 41: Propensity Score Assessment Plot for Logistic Regression Stratification: Grade 8 Math

Table 21: Logistic Regression Stratification Results for Grade 8 math

	Public		Charter	
Strata	Mean	n	Mean	n
1	278.58	7402	273.03	135
2	289.98	7359	288.65	177
3	294.62	7334	292.99	202
4	298.77	7292	297.52	244
5	294.60	7207	297.14	329
6	277.44	7183	275.17	353
7	270.63	7132	267.17	404
8	263.73	7022	264.47	514
9	259.42	6864	257.97	672
10	262.71	6733	261.67	803

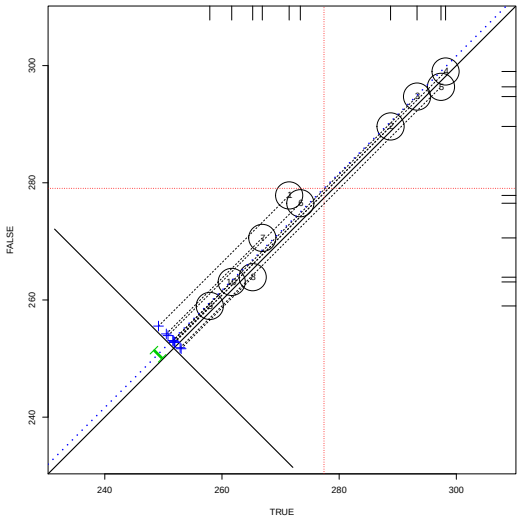


Figure 42: Propensity Score Assessment Plot for Logistic Regression AIC Stratification: Grade 8 Math

Table 22: Logistic Regression AIC Stratification Results for Grade 8 math

	Public		Charter	
Strata	Mean	n	Mean	n
1	277.84	7407	271.47	135
2	289.62	7388	288.77	172
3	294.71	7324	293.29	211
4	299.00	7268	298.16	240
5	296.38	7199	297.38	338
6	276.51	7193	273.37	342
7	270.59	7122	266.88	417
8	263.89	7033	265.24	512
9	258.98	6862	257.92	666
10	263.08	6732	261.66	800

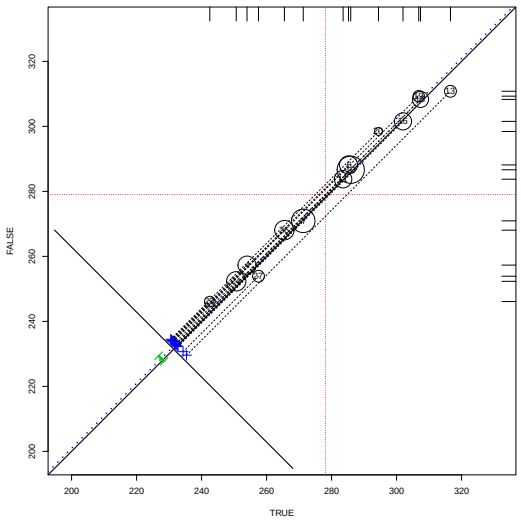


Figure 43: Propensity Score Assessment Plot for Classification Tree Stratification: Grade 8 Math

Table 23: Classification Trees Stratification Results for Grade 8 math

	Public		Charter	
Strata	Mean	n	Mean	n
4	257.31	5529	253.98	261
6	288.14	5544	285.21	257
7	270.93	10414	271.28	704
10	286.65	15771	285.89	318
11	283.75	5109	283.55	156
13	310.81	2000	316.58	51
16	301.54	4895	301.98	166
18	308.30	4163	307.37	172
19	309.29	1720	306.81	106
20	298.44	694	294.44	48
23	246.10	1432	242.53	92
24	252.32	6261	250.66	581
26	268.06	6244	265.49	670
27	253.90	1752	257.52	251

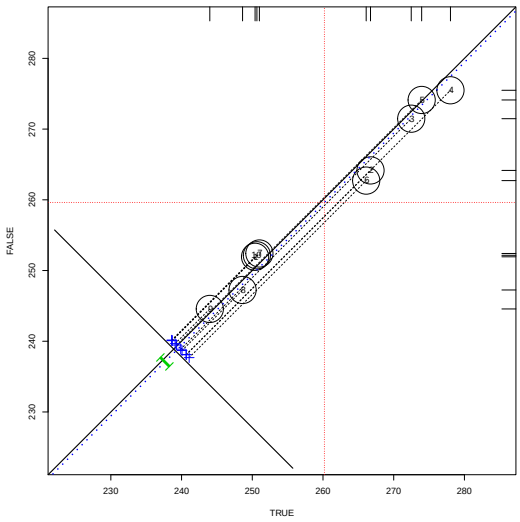


Figure 44: Propensity Score Assessment Plot for Logistic Regression Stratification: Grade 8 Reading

Table 24: Logistic Regression Stratification Results for Grade 8 read

	Public		Charter	
Strata	Mean	n	Mean	n
1	251.94	7634	250.40	129
2	264.16	7596	266.72	167
3	271.47	7535	272.48	226
4	275.49	7519	278.03	243
5	274.12	7471	273.95	291
6	262.72	7413	266.11	349
7	252.41	7379	251.00	383
8	247.25	7213	248.63	549
9	244.57	7103	243.99	659
10	252.15	6947	250.66	816

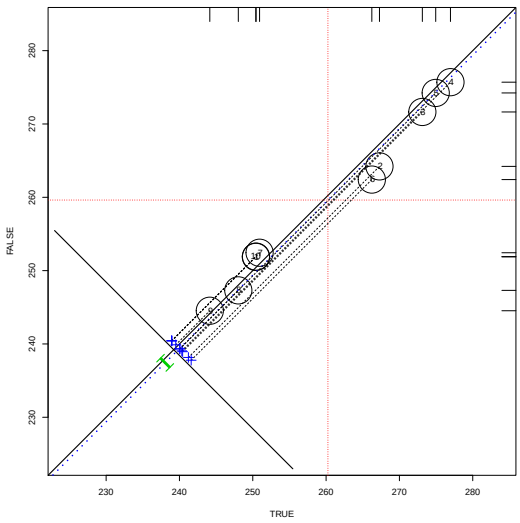


Figure 45: Propensity Score Assessment Plot for Logistic Regression AIC Stratification: Grade 8 Reading

Table 25: Logistic Regression AIC Stratification Results for Grade 8 read

	Public		Charter	
Strata	Mean	n	Mean	n
1	251.87	7636	250.42	128
2	264.22	7592	267.28	169
3	271.64	7530	273.13	232
4	275.70	7530	276.96	232
5	274.23	7466	274.95	296
6	262.42	7417	266.24	345
7	252.44	7365	250.93	397
8	247.31	7208	248.03	554
9	244.53	7132	244.14	630
10	251.91	6934	250.44	829

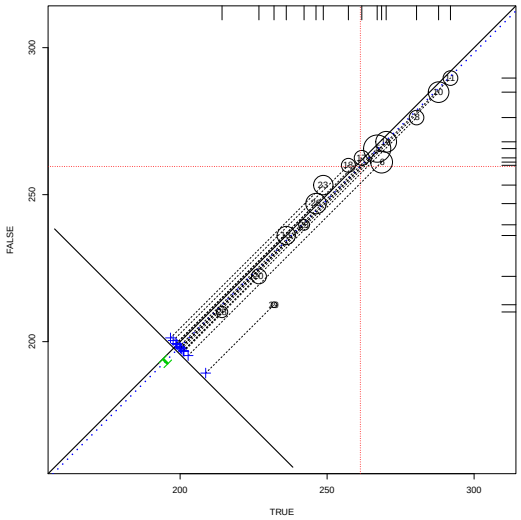


Figure 46: Propensity Score Assessment Plot for Classification Tree Stratification: Grade 8 Reading

Table 26: Classification Trees Stratification Results for Grade 8 read

	Public		Charter	
Strata	Mean	n	Mean	n
5	265.67	14613	267.03	381
6	261.08	8482	268.53	163
8	276.22	3089	280.42	79
10	284.87	7238	287.92	265
11	289.66	3063	291.96	157
13	236.11	5341	236.11	233
17	262.51	3239	261.80	188
18	259.98	2725	257.24	112
19	267.97	7196	270.08	453
20	222.21	3127	226.81	250
23	253.26	5887	248.70	529
25	239.75	1400	242.16	119
26	246.96	6535	246.22	766
28	210.10	1606	214.26	84
29	212.56	269	231.94	33

Appendix G

Multilevel PSA Covariate Balance Plots

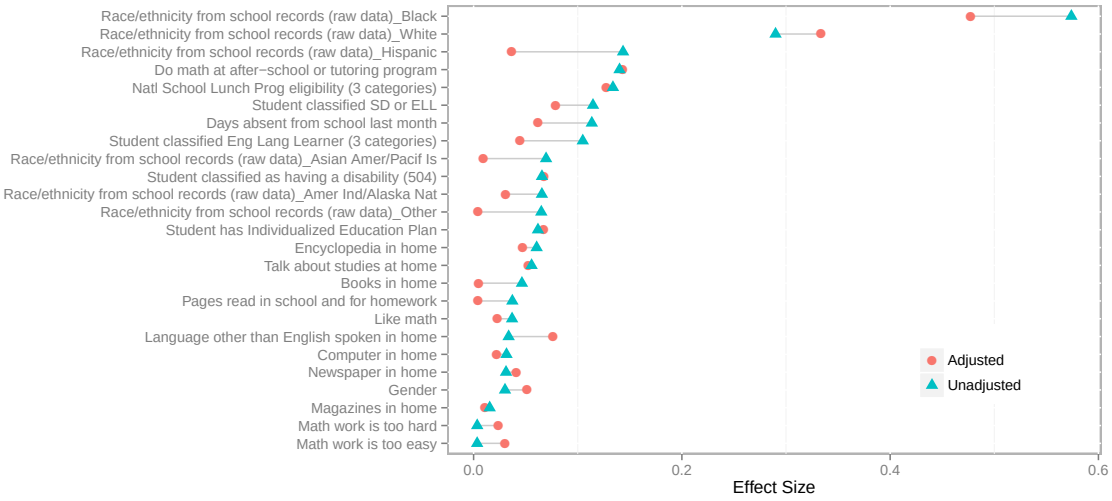


Figure 47: Multilevel PSA Covariate Balance Plot Logistic Regression: Grade 4 Math

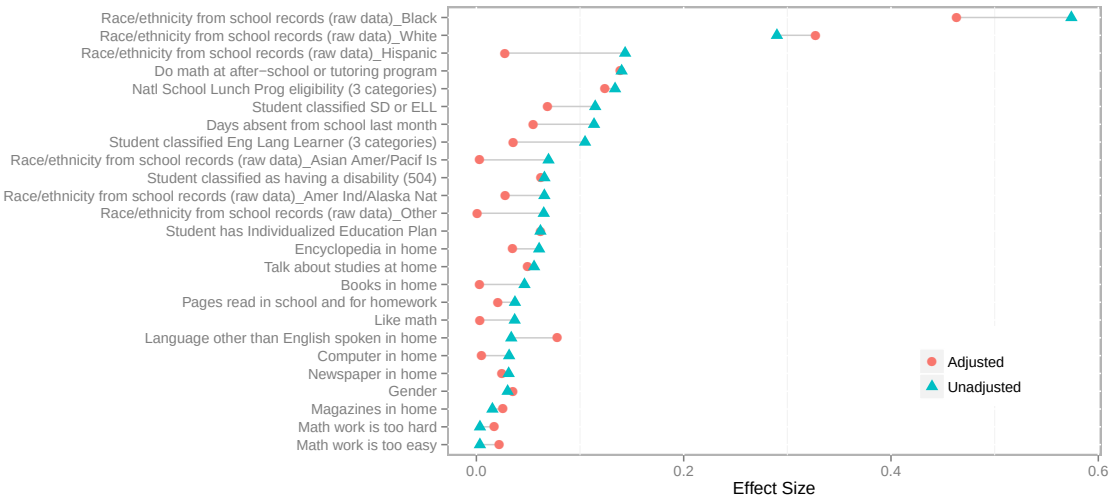


Figure 48: Multilevel PSA Covariate Balance Plot Logistic Regression AIC: Grade 4 Math

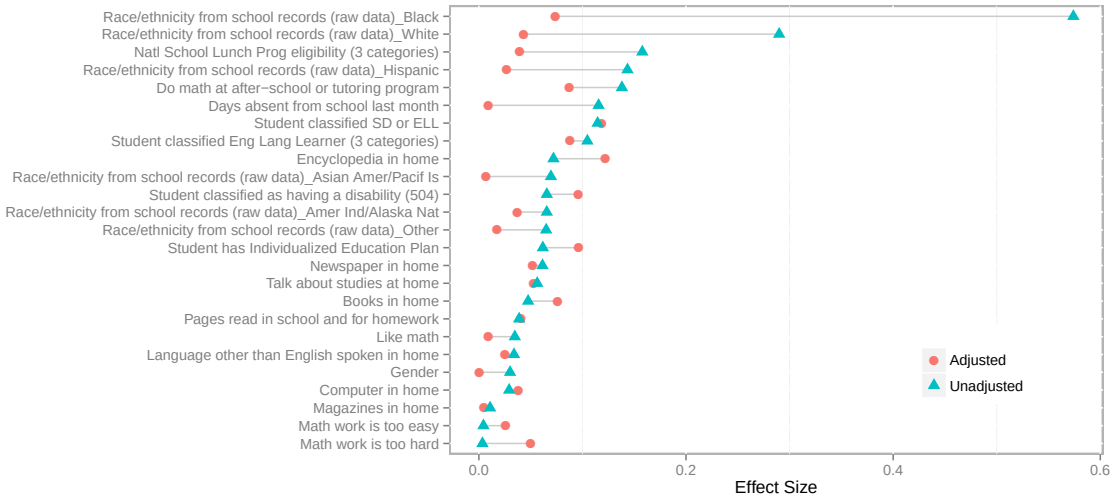


Figure 49: Multilevel PSA Covariate Balance Plot Classification Tree: Grade 4 Math

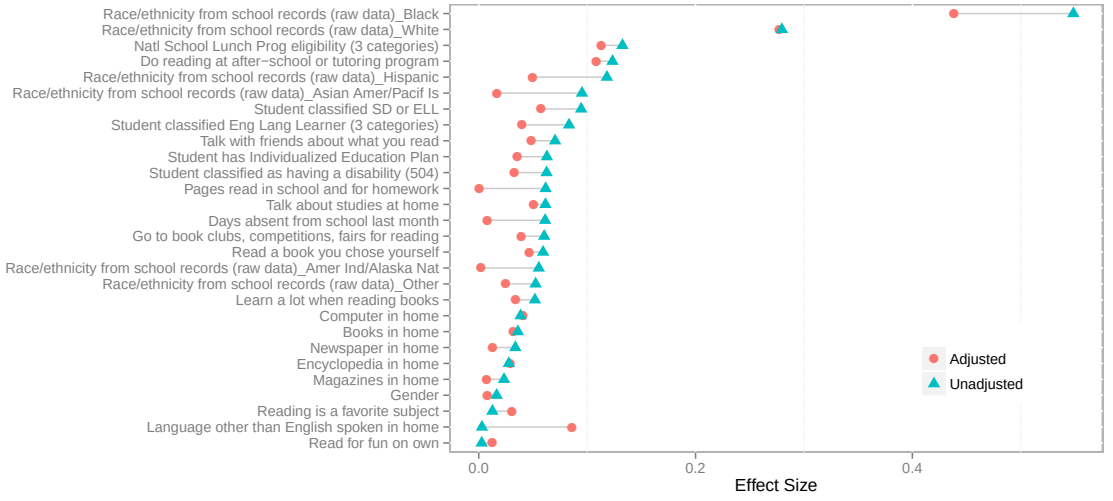


Figure 50: Multilevel PSA Covariate Balance Plot Logistic Regression: Grade 4 Reading

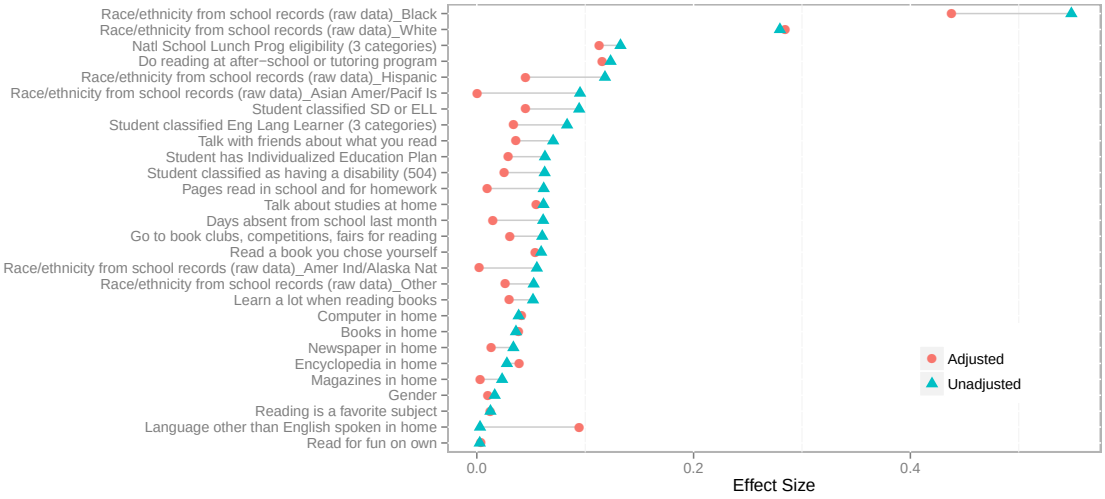


Figure 51: Multilevel PSA Covariate Balance Plot Logistic Regression AIC: Grade 4 Reading

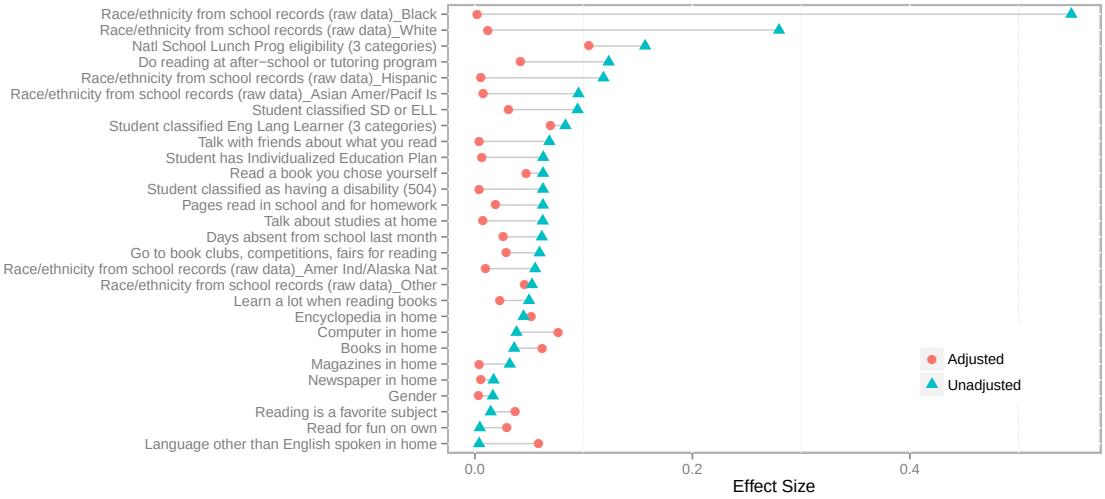


Figure 52: Multilevel PSA Covariate Balance Plot Classification Tree: Grade 4 Reading

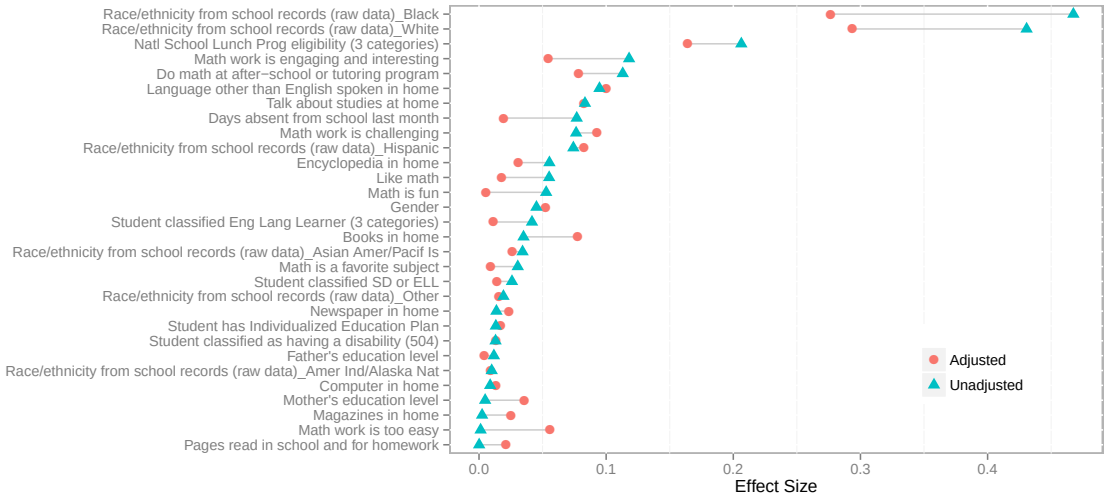


Figure 53: Multilevel PSA Covariate Balance Plot Logistic Regression: Grade 8 Math

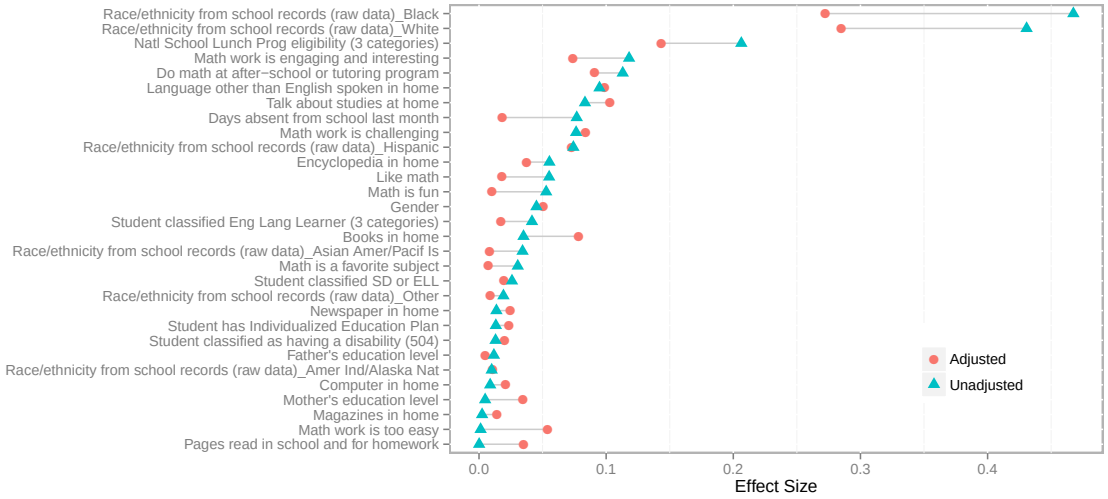


Figure 54: Multilevel PSA Covariate Balance Plot Logistic Regression AIC: Grade 8 Math

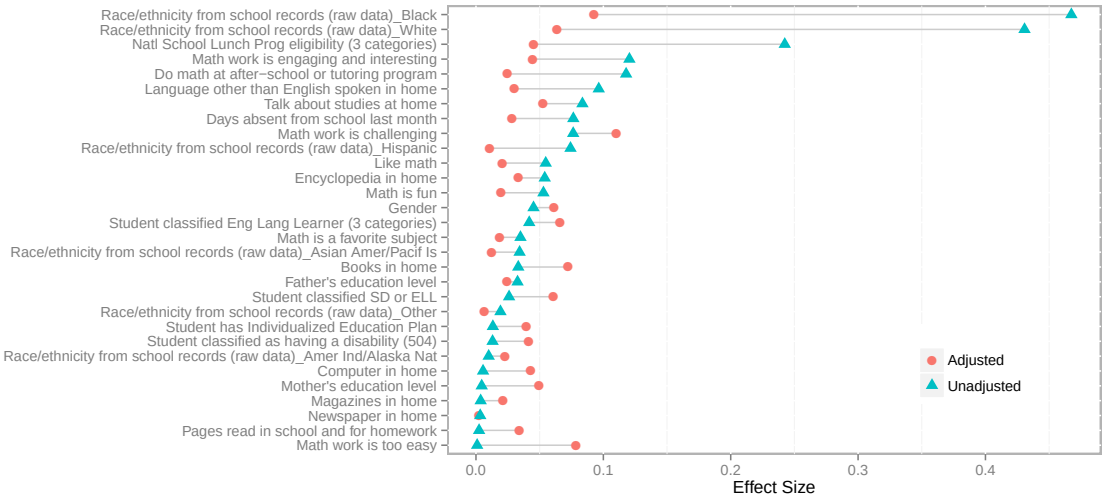


Figure 55: Multilevel PSA Covariate Balance Plot Classification Tree: Grade 8 Math

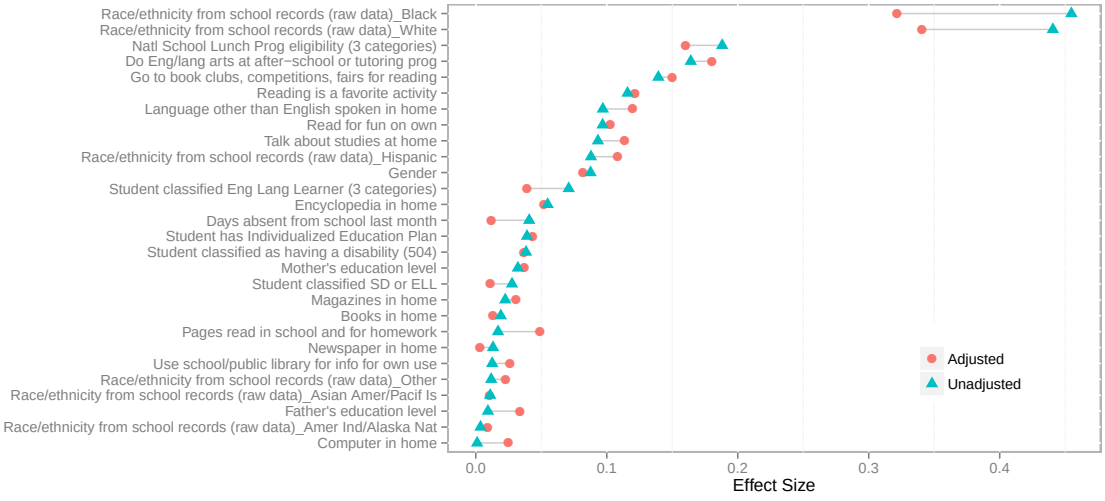


Figure 56: Multilevel PSA Covariate Balance Plot Logistic Regression: Grade 8 Reading

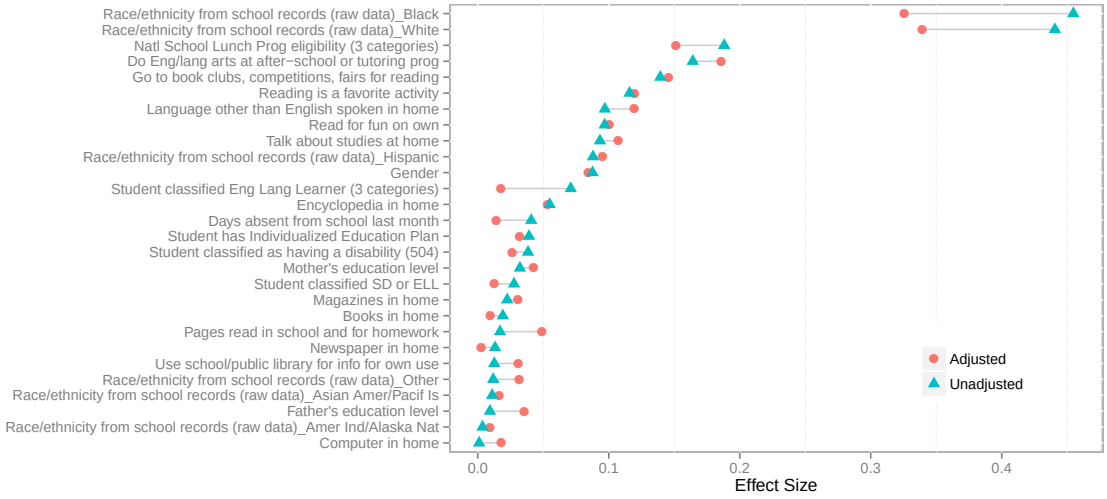


Figure 57: Multilevel PSA Covariate Balance Plot Logistic Regression AIC: Grade 8 Reading

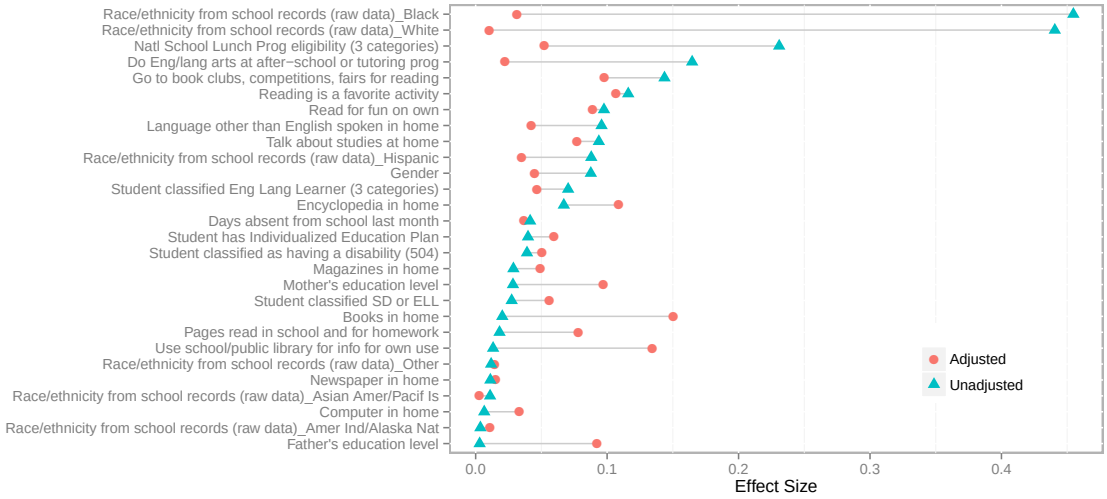


Figure 58: Multilevel PSA Covariate Balance Plot Classification Tree: Grade 8 Reading

Appendix H

Multilevel PSA Results

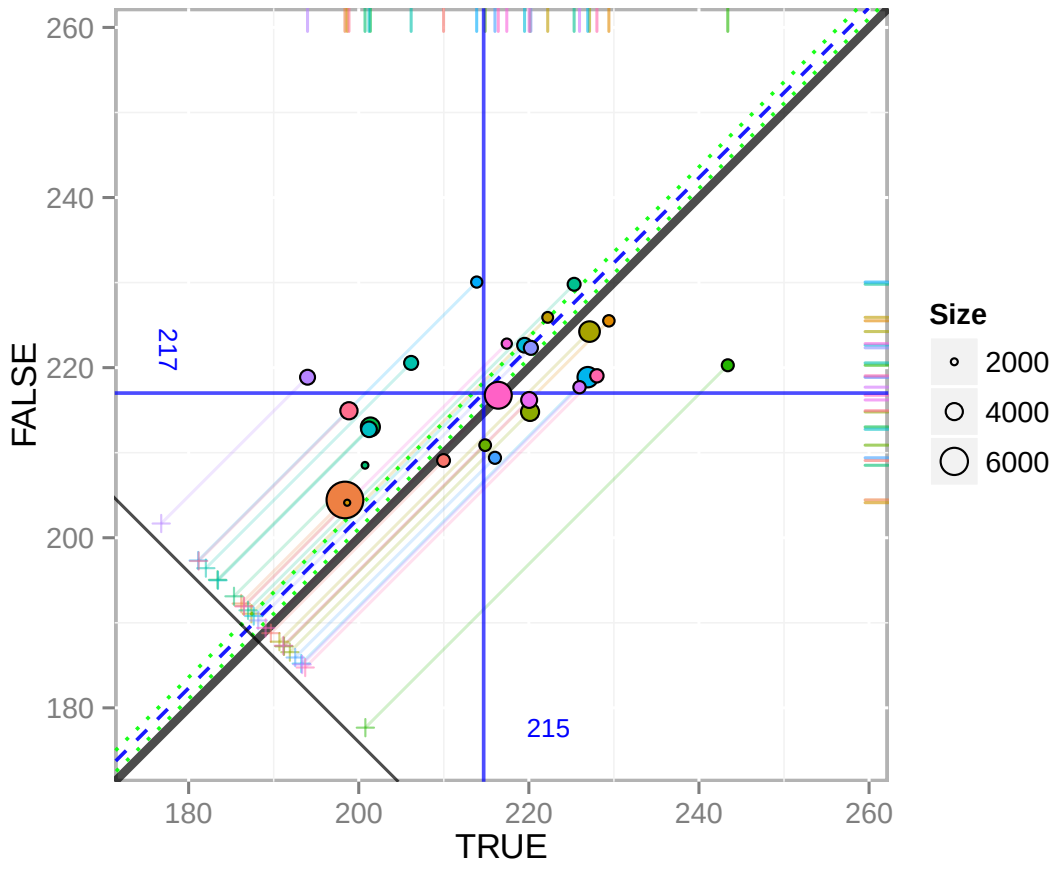


Figure 59: Multilevel PSA Assessment Plot Logistic Regression: Grade 4 Reading

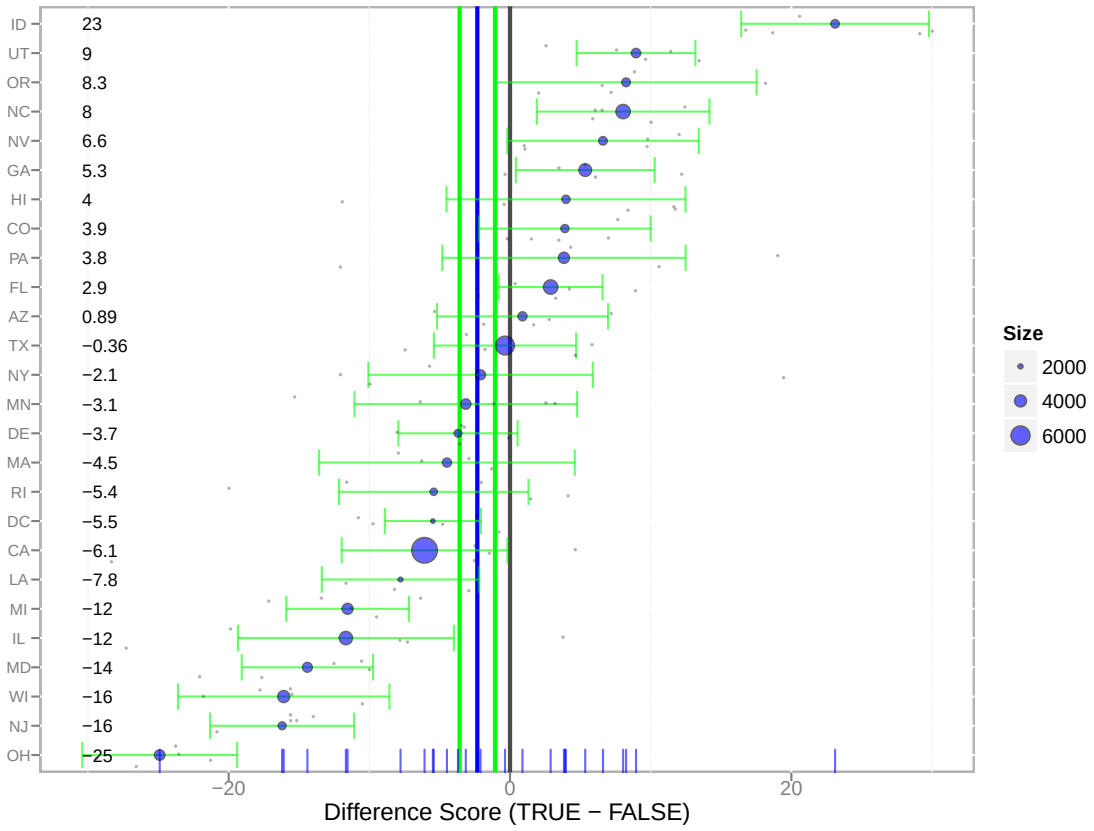


Figure 60: Multilevel PSA Difference Plot Logistic Regression: Grade 4 Reading

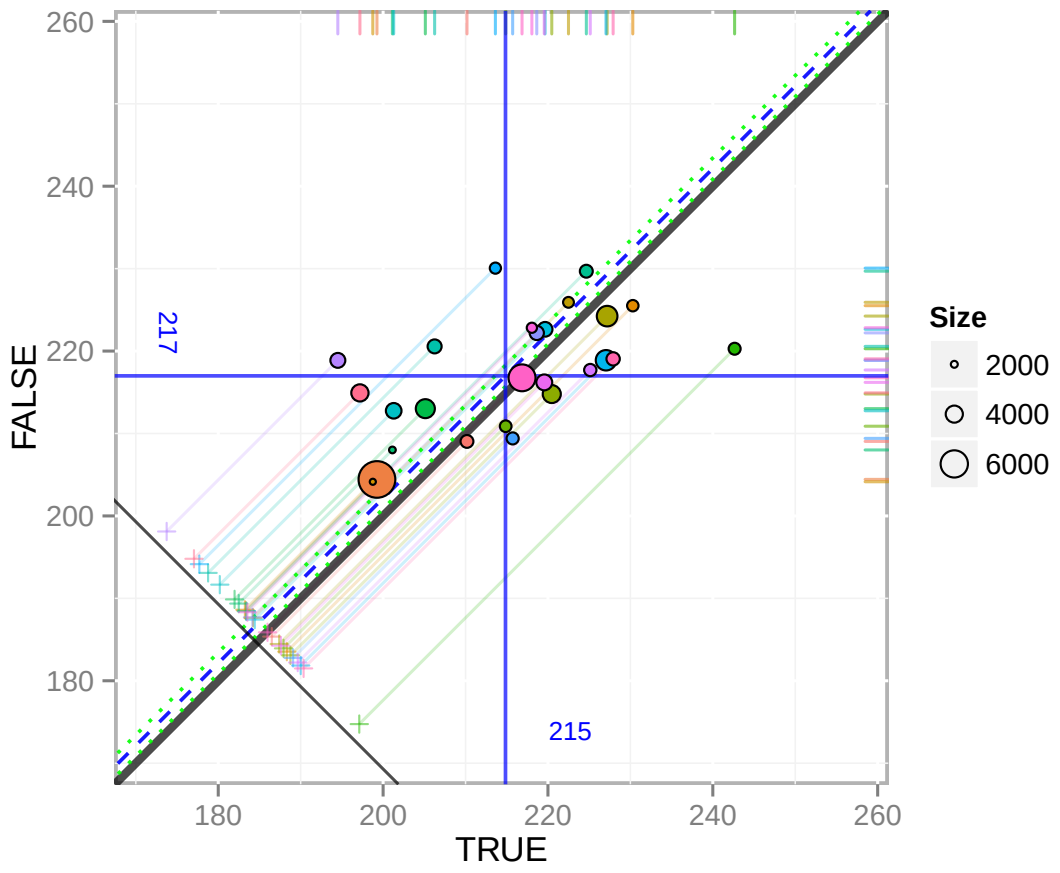


Figure 61: Multilevel PSA Assessment Plot Logistic Regression AIC: Grade 4 Reading

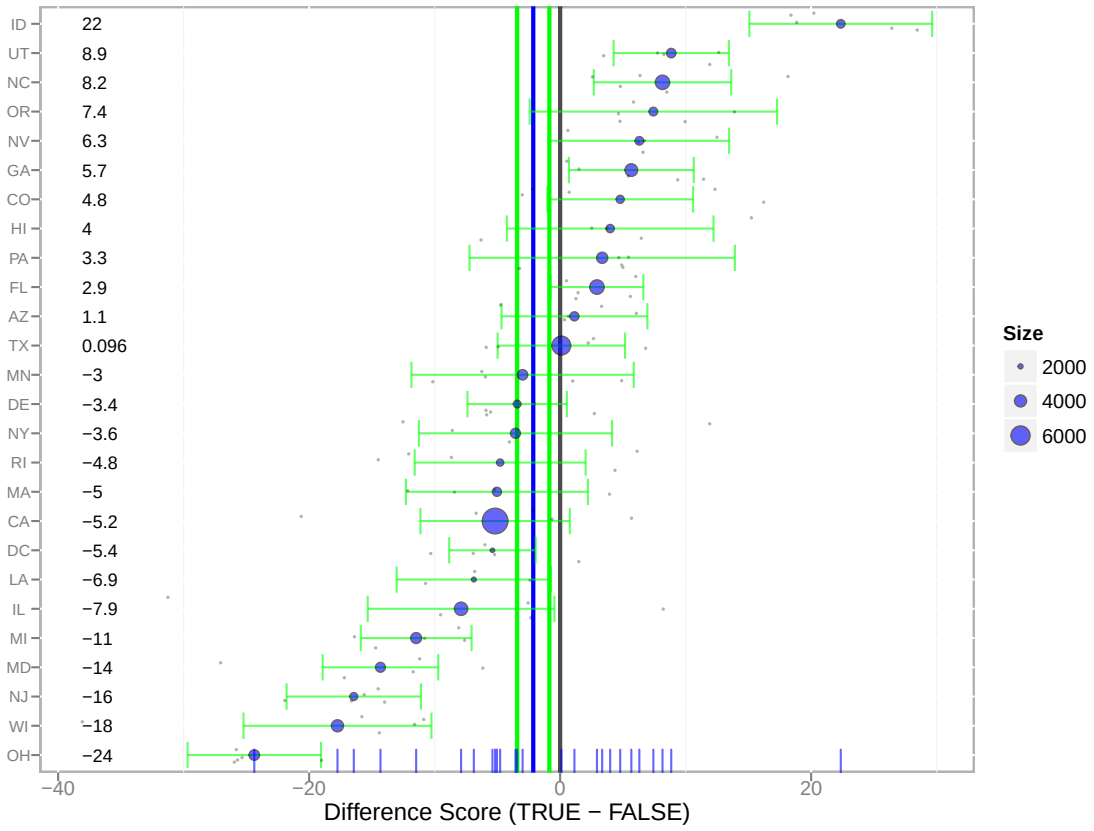


Figure 62: Multilevel PSA Difference Plot Logistic Regression AIC: Grade 4 Reading

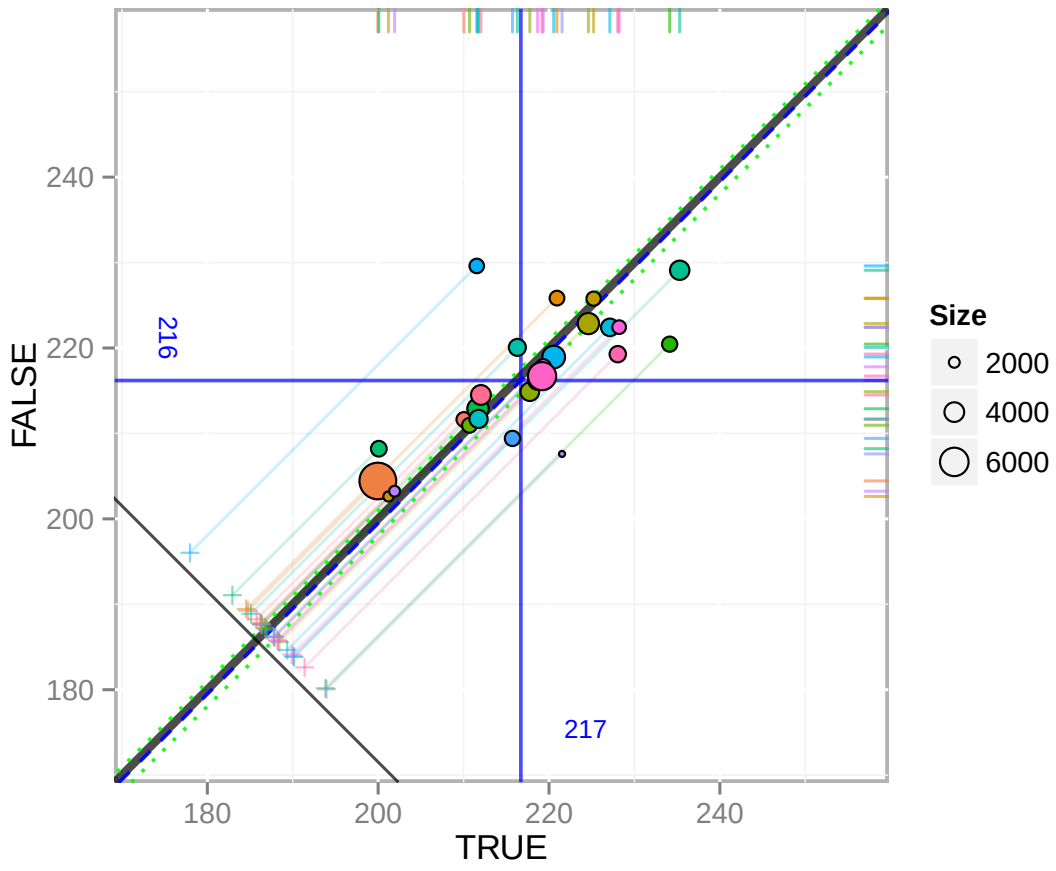


Figure 63: Multilevel PSA Assessment Plot Classification Trees: Grade 4 Reading

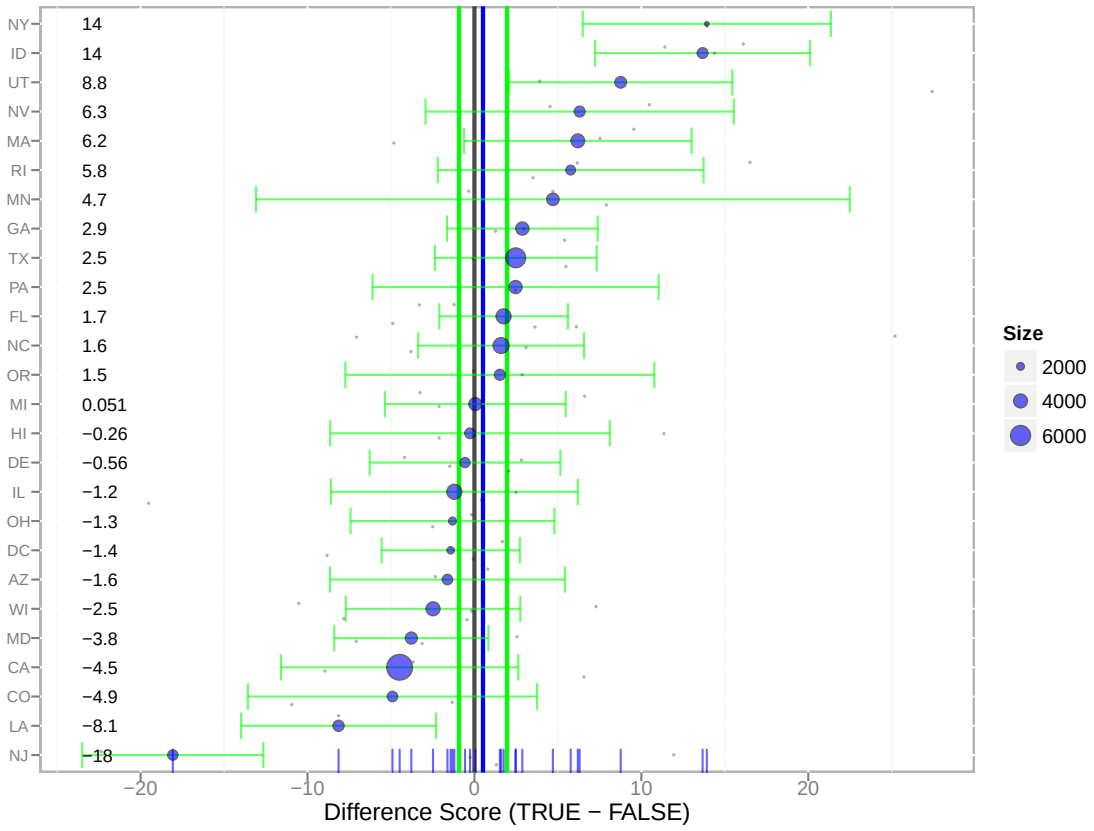


Figure 64: Multilevel PSA Difference Plot Classification Trees: Grade 4 Reading

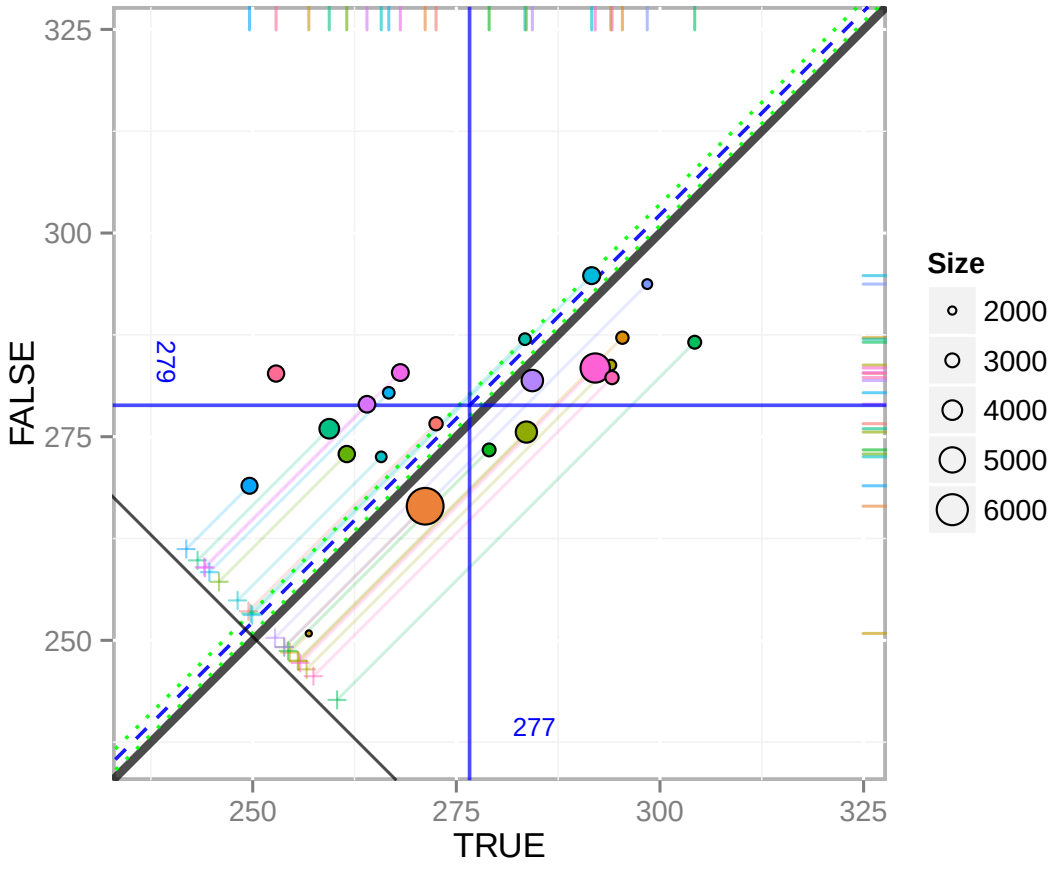


Figure 65: Multilevel PSA Assessment Plot Logistic Regression: Grade 8 Math

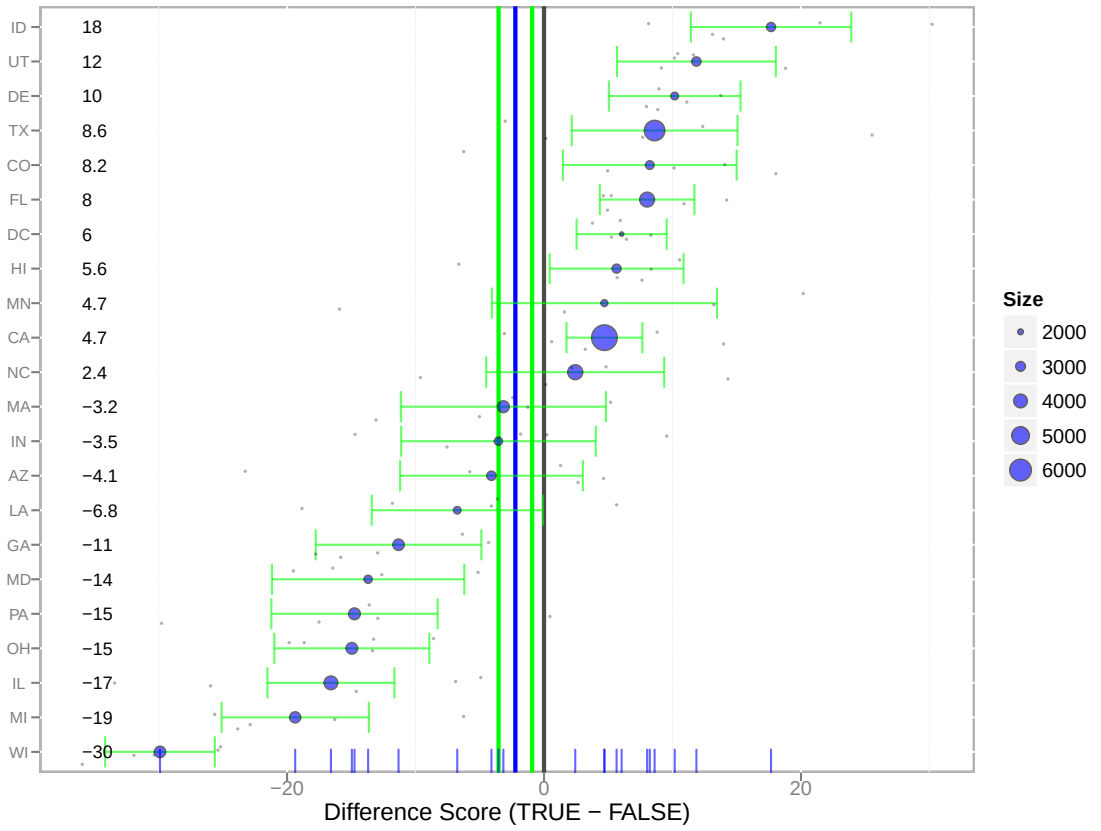


Figure 66: Multilevel PSA Difference Plot Logistic Regression: Grade 8 Math

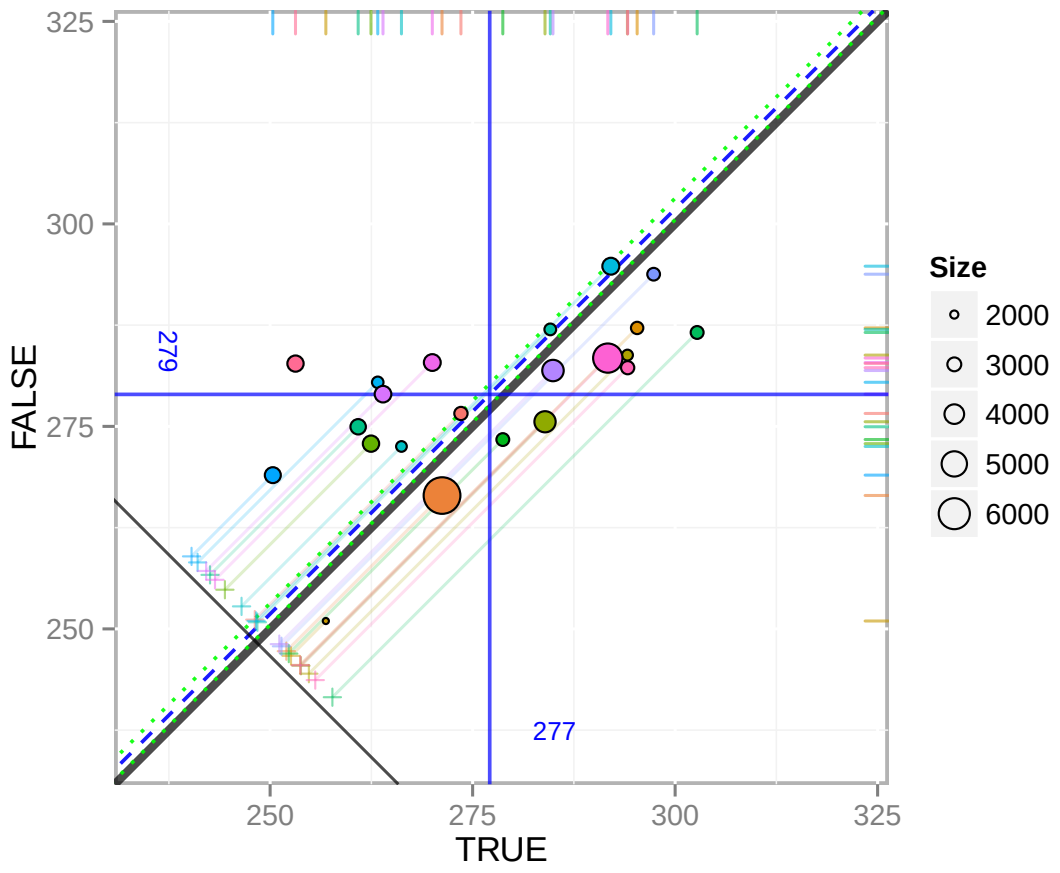


Figure 67: Multilevel PSA Assessment Plot Logistic Regression AIC: Grade 8 Math

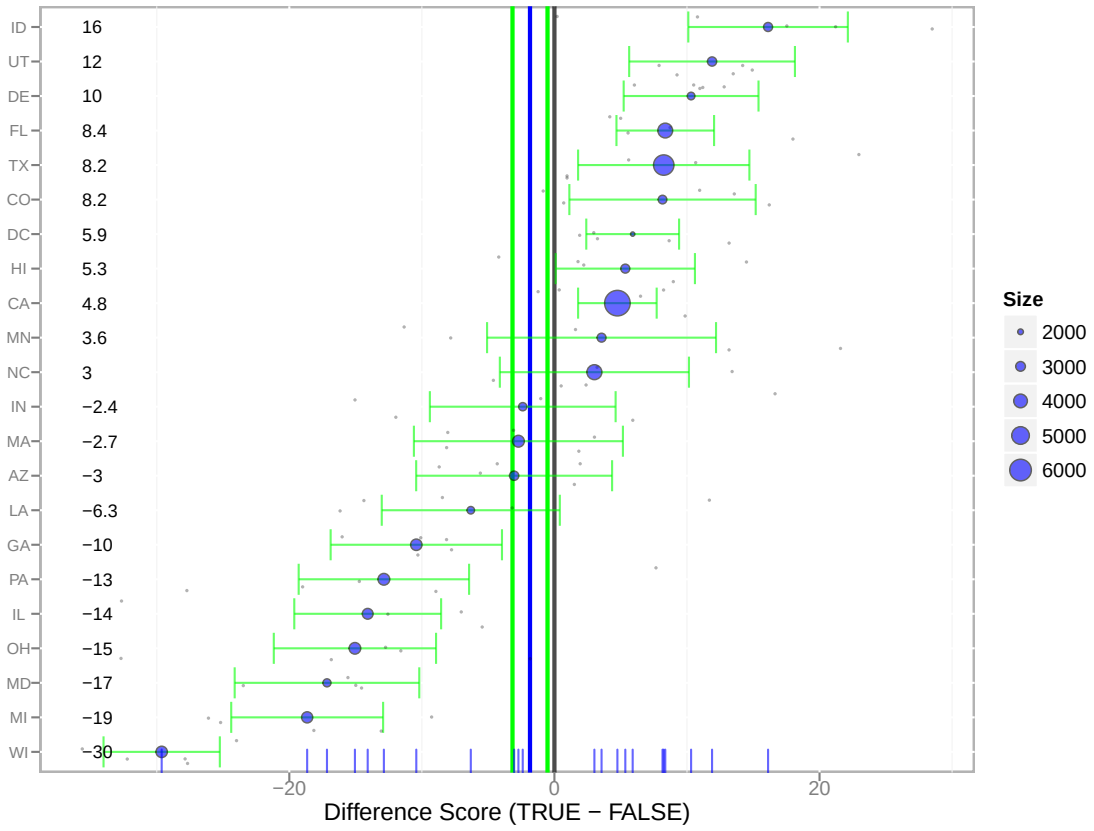


Figure 68: Multilevel PSA Difference Plot Logistic Regression AIC: Grade 8 Math

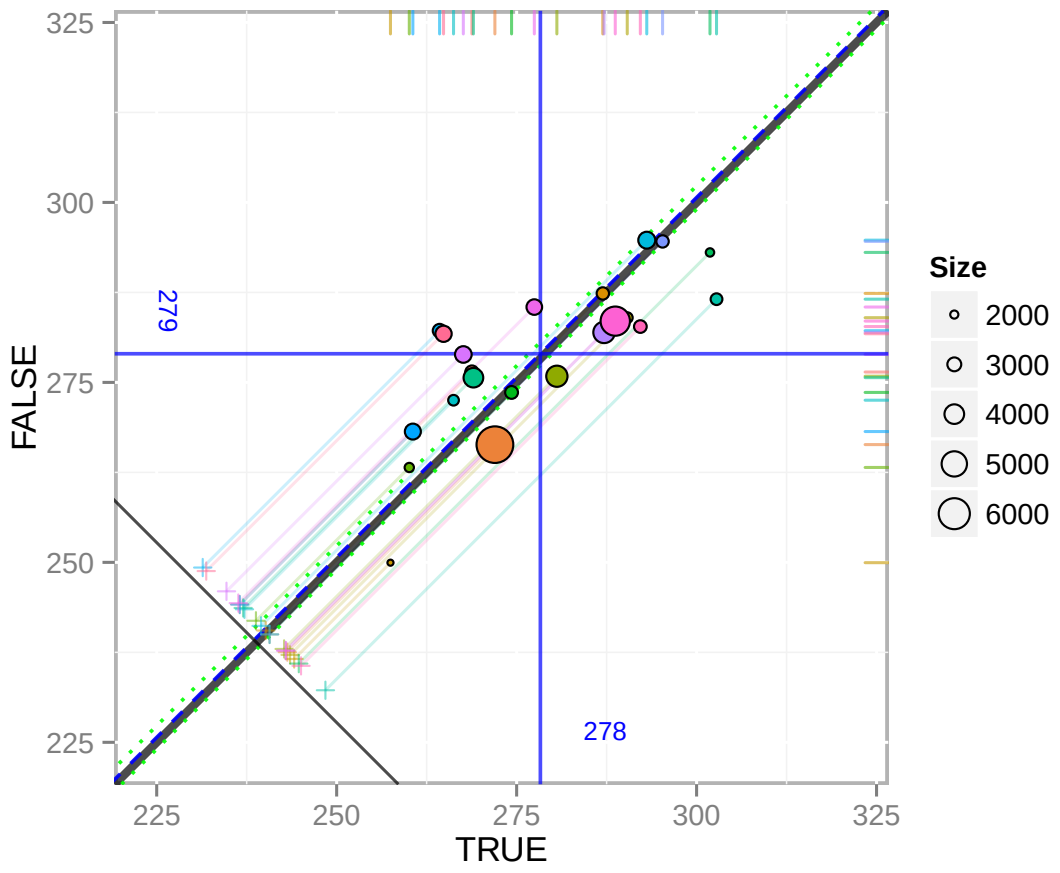


Figure 69: Multilevel PSA Assessment Plot Classification Trees: Grade 8 Math

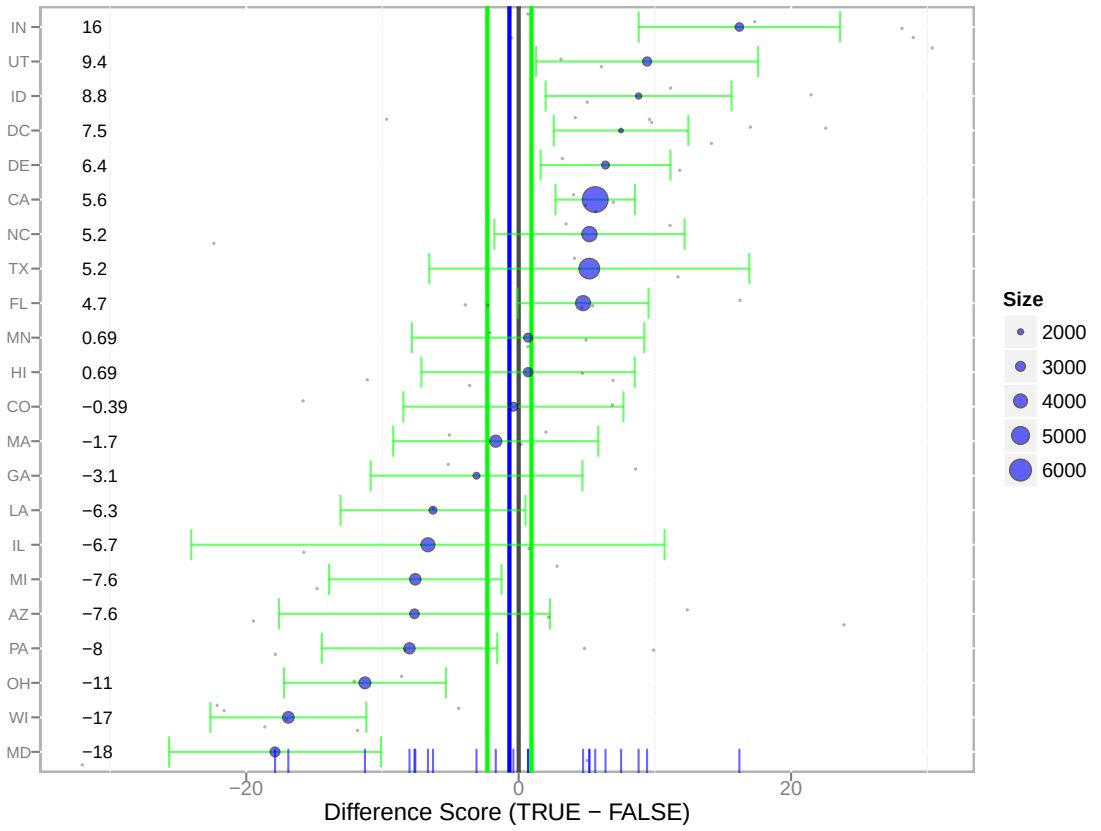


Figure 70: Multilevel PSA Difference Plot Classification Trees: Grade 8 Math

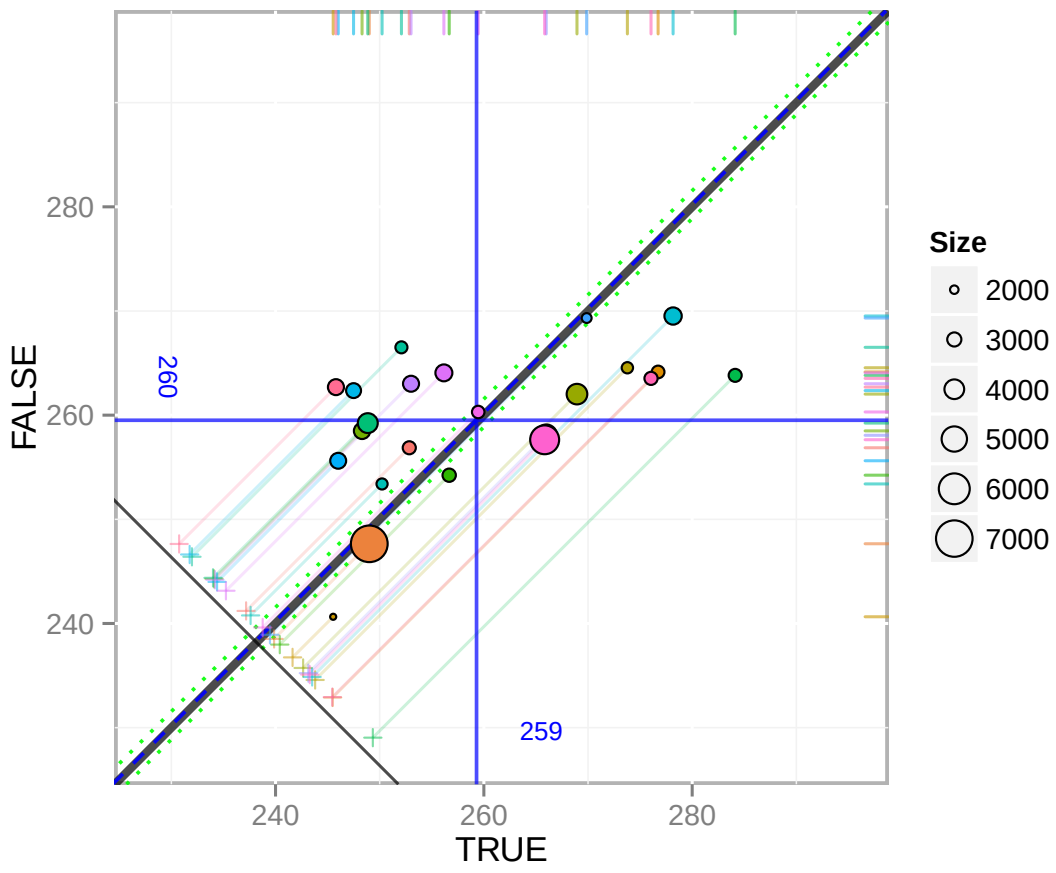


Figure 71: Multilevel PSA Assessment Plot Logistic Regression: Grade 8 Reading

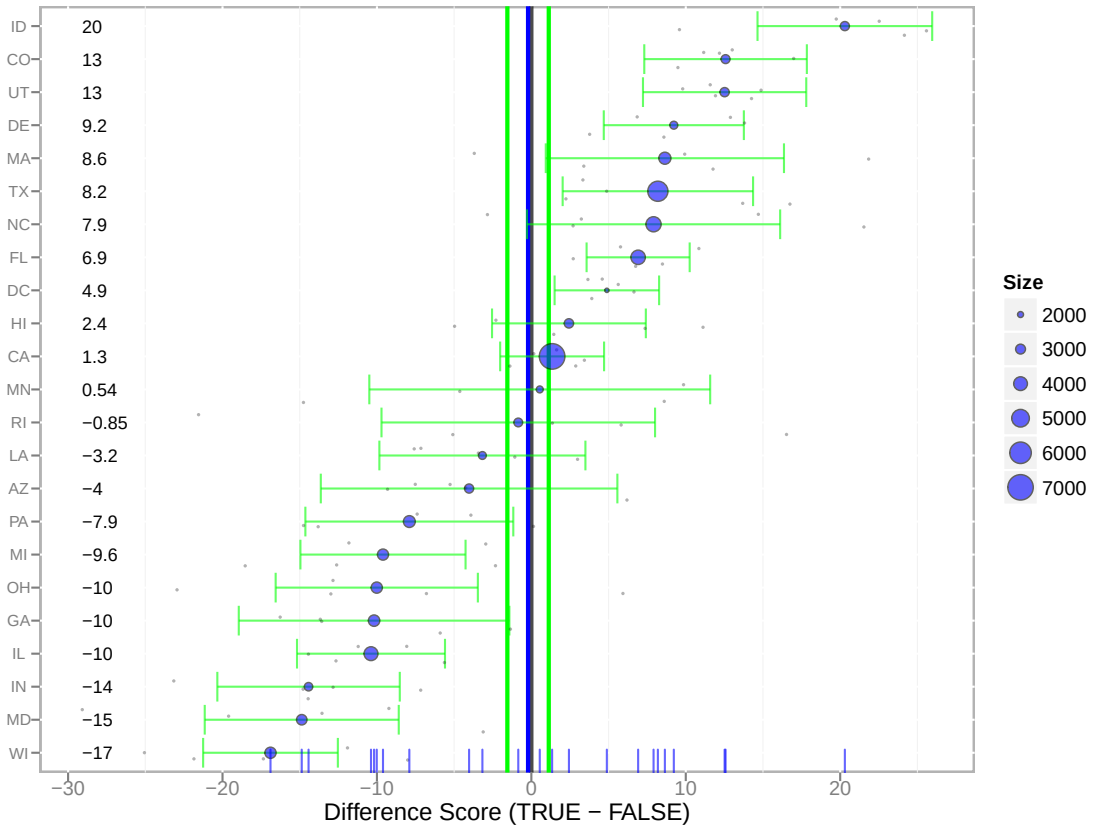


Figure 72: Multilevel PSA Difference Plot Logistic Regression: Grade 8 Reading

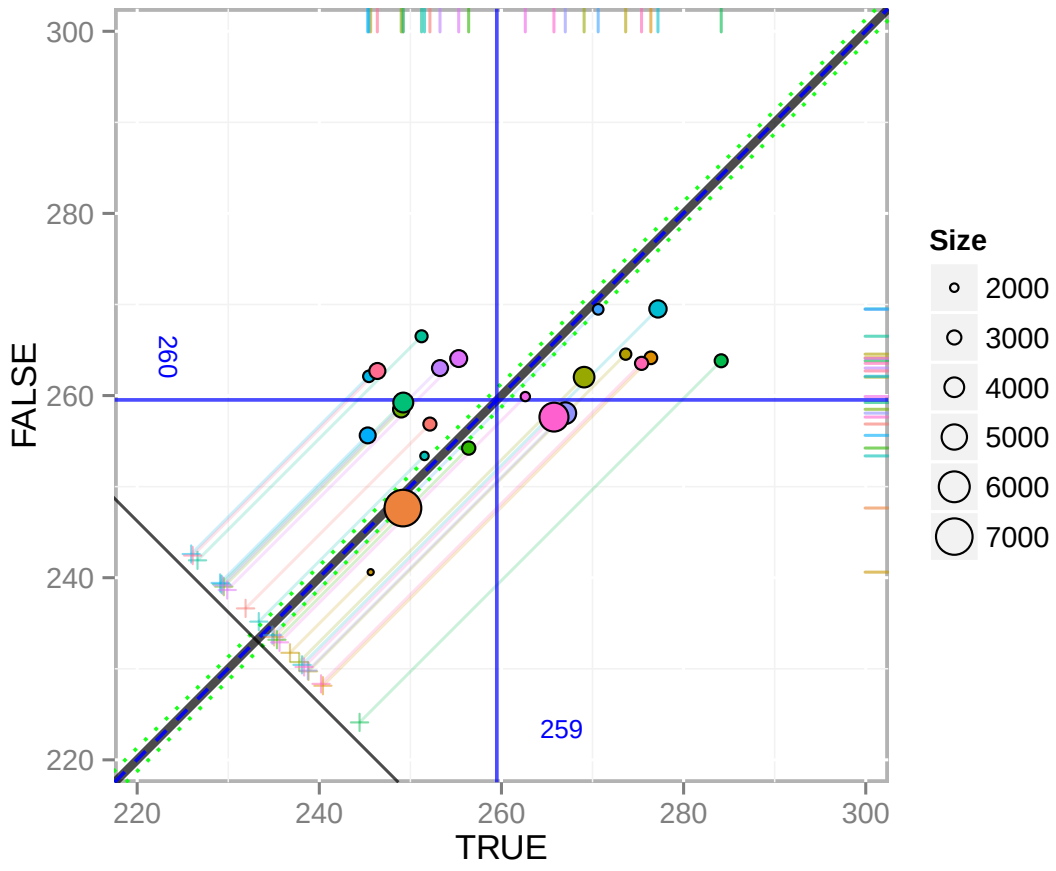


Figure 73: Multilevel PSA Assessment Plot Logistic Regression AIC: Grade 8 Reading

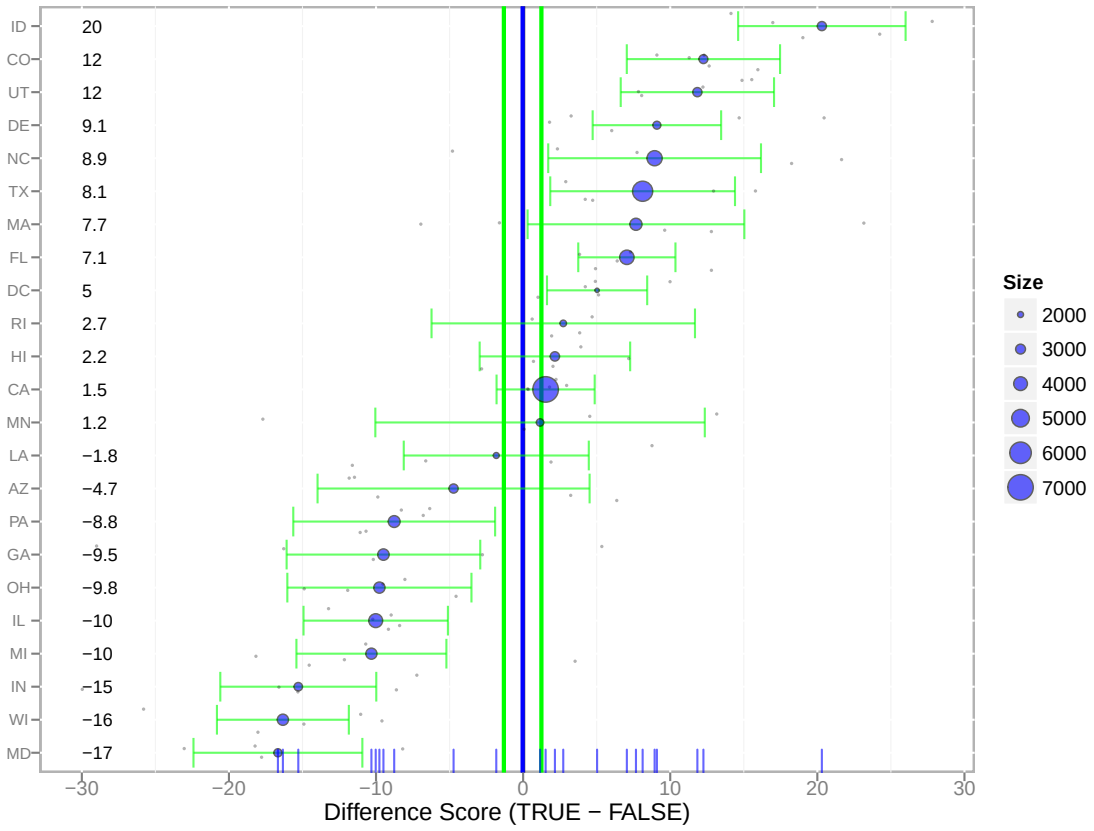


Figure 74: Multilevel PSA Difference Plot Logistic Regression AIC: Grade 8 Reading

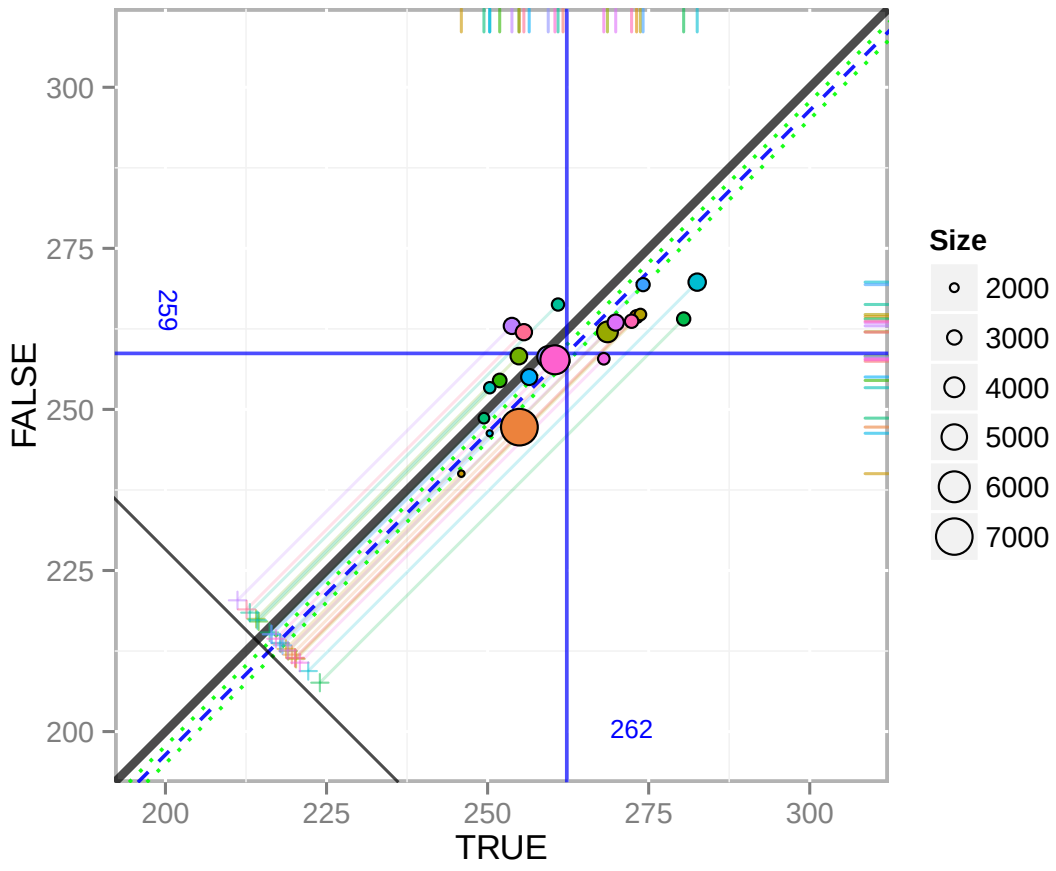


Figure 75: Multilevel PSA Assessment Plot Classification Trees: Grade 8 Reading

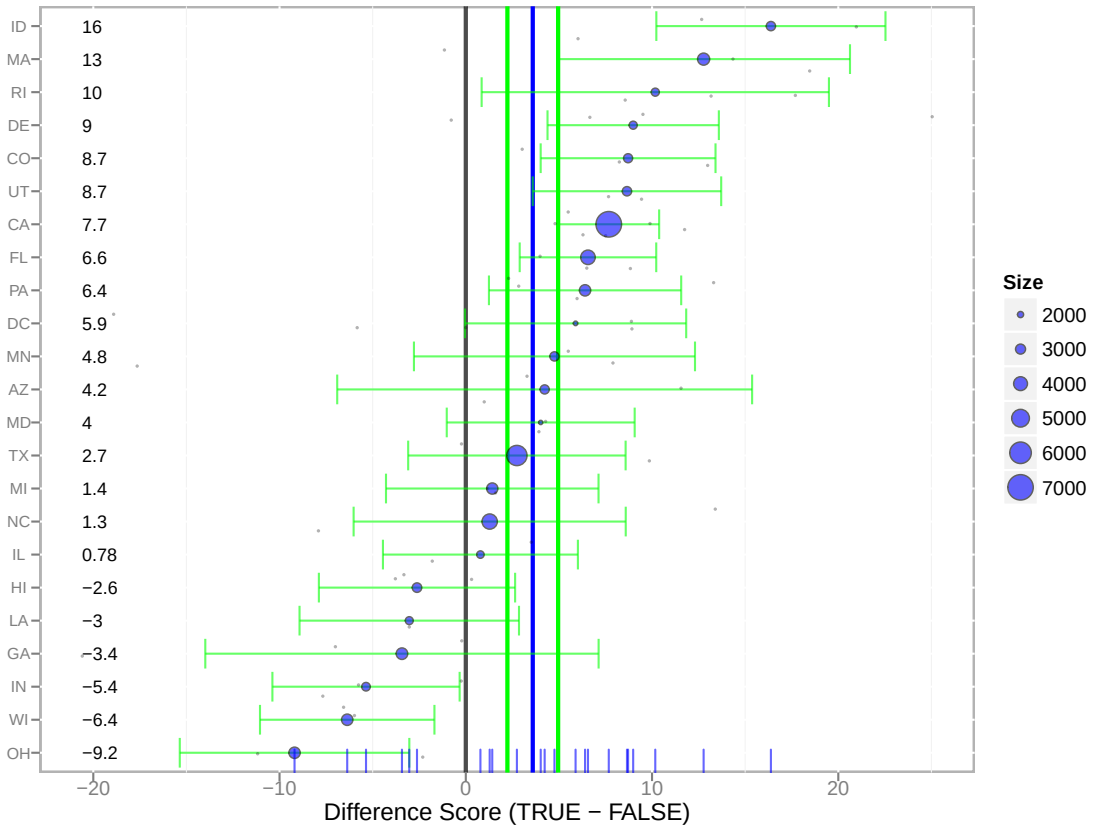


Figure 76: Multilevel PSA Difference Plot Classification Trees: Grade 8 Reading

Appendix I

Multilevel PSA Classification Tree Heat Maps

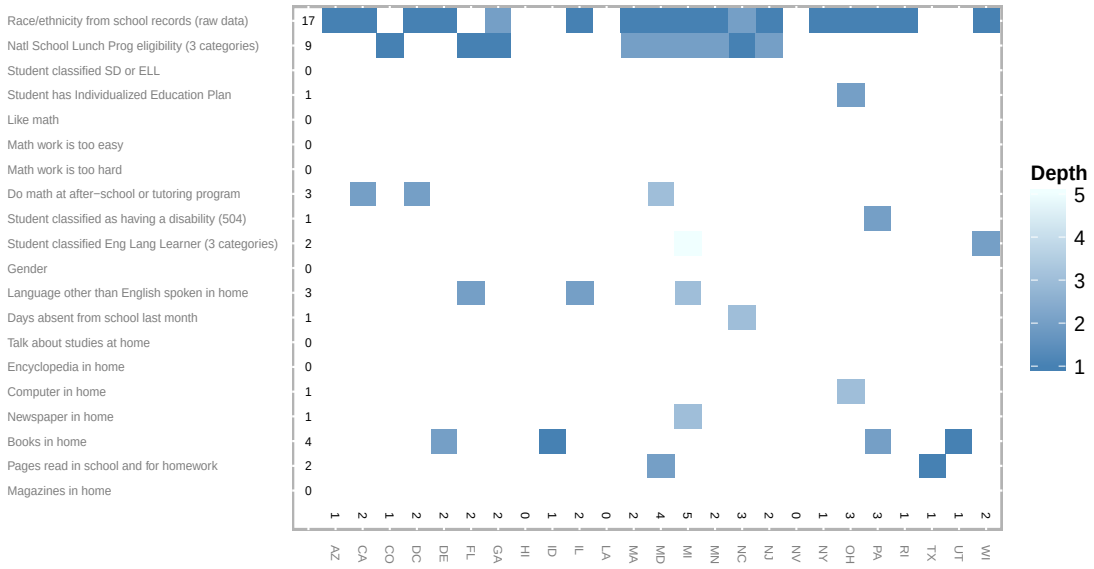


Figure 77: Heat Map of Relative Importance of Covariates from Classification Trees: Grade 4 Math

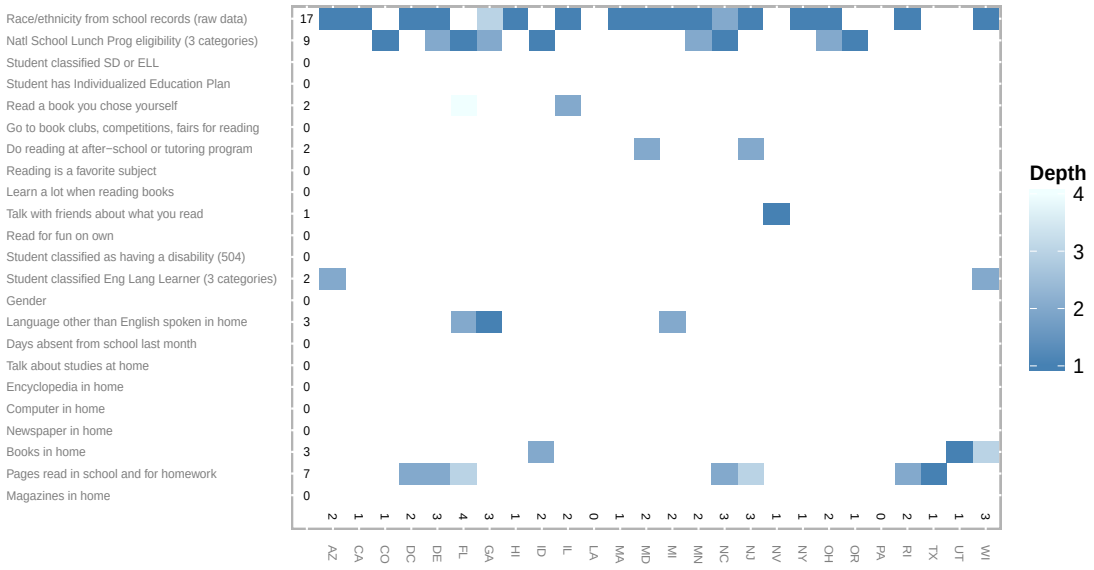


Figure 78: Heat Map of Relative Importance of Covariates from Classification Trees: Grade 4 Reading

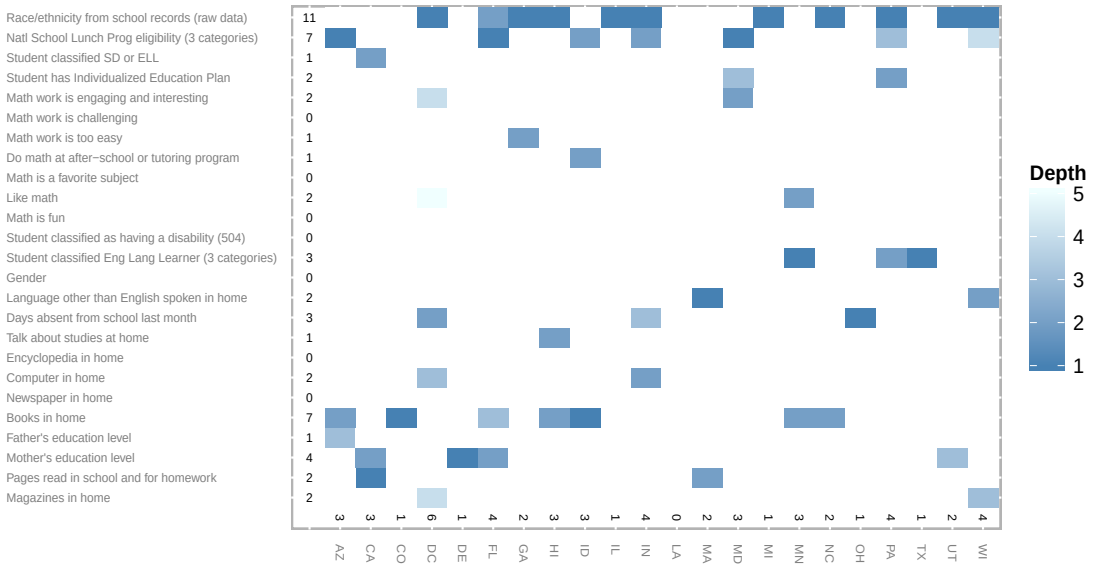


Figure 79: Heat Map of Relative Importance of Covariates from Classification Trees: Grade 8 Math

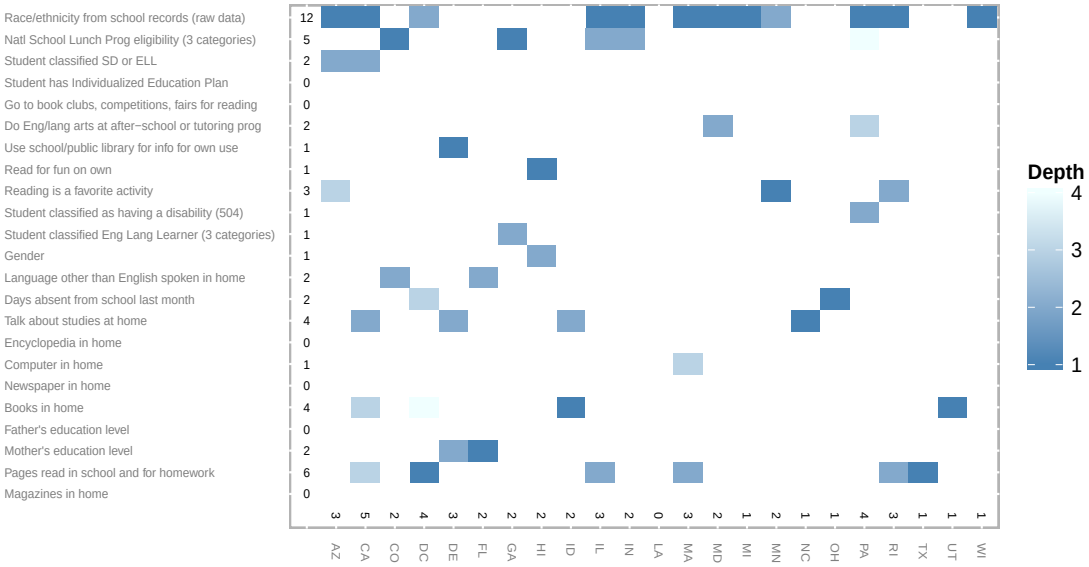


Figure 80: Heat Map of Relative Importance of Covariates from Classification Trees: Grade 8 Reading

Appendix J
multilevelPSA R Package

The `multilevelPSA` R package was developed, in part, to conduct the analysis for this dissertation. It is available from the Comprehensive R Archive Network (CRAN) at `http://cran.r-project.org/web/packages/multilevelPSA`. The latest version can be installed using the `install.packages` function in R:

```
> install.packages('multilevelPSA', repos='http://cran.r-project.org')
```

The following list provides brief descriptions of the key functions in the `multilevelPSA` package. More information is available vis-à-vis the R help system.

`getPropensityScores` Returns a data frame with two columns corresponding to the level 2 variable and the fitted value from the logistic regression.

`getStrata` Returns a data frame with two columns corresponding to the level 2 variable and the leaves from the conditional inference trees.

`loess.plot` Loess plot with density distributions for propensity scores and outcomes on top and right, respectively.

`missing.plot` Returns a heat map graphic representing missingness of variables grouped by the given grouping vector.

`mlpsa` This function will perform phase II of the multilevel propensity score analysis.

`plot.mlpsa` Creates the multilevel assessment plot.

`mlpsa.circ.plot` Plots the results of a multilevel propensity score model.

`mlpsa.ctree` Estimates propensity scores using the recursive partitioning in a conditional inference framework.

`mlpsa.difference.plot` Creates a graphic summarizing the differences between treatment and comparison groups within and across level two clusters.

`mlpsa.distribution.plot` Plots distribution for either the treatment or comparison group.

`mlpsa.logistic` Estimates propensity scores using logistic regression.

`psrange` Estimates models with increasing number of comparison subjects starting from 1:1 to using all available comparison group subjects.

`plot.psrange` Plots the results of `psrange`.

`tree.plot` Heat map representing variables used in a conditional inference tree across level 2 variables.

Appendix K

Simulating Propensity Score Ranges

The `getSimulatedData` function is what is used to simulate covariates with varying overlap.

```
getSimulatedData <- function(nvars = 3,
    ntreat = 100, treat.mean = 0.6, treat.sd = 0.5,
    ncontrol = 1000, control.mean = 0.4, control.sd = 0.5) {
    if (length(treat.mean) == 1) {
        treat.mean = rep(treat.mean, nvars)
    }
    if (length(treat.sd) == 1) {
        treat.sd = rep(treat.sd, nvars)
    }
    if (length(control.mean) == 1) {
        control.mean = rep(control.mean, nvars)
    }
    if (length(control.sd) == 1) {
        control.sd = rep(control.sd, nvars)
    }

    df <- c(rep(0, ncontrol), rep(1, ntreat))
    for (i in 1:nvars) {
        df <- cbind(df, c(
            rnorm(ncontrol, mean = control.mean[1], sd = control.sd[1]),
            rnorm(ntreat, mean = treat.mean[1], sd = treat.sd[1])
        ) )
    }
    df <- as.data.frame(df)
    names(df) <- c("treat", letters[1:nvars])
    return(df)
}
```

The following code was used to create Figure 14 in Chapter 5 which represents moderate overlap but some separation in covariates between treatment and control.

```
set.seed(2112)
df.psrange <- getSimulatedData(ncontrol = 1000, nvars=1,
                               treat.mean=0.6, treat.sd=0.4,
                               control.mean=0.4, control.sd=0.4)
psrange.test <- psrange(df.psrange, df.psrange$treat, treat ~ .,
                        samples = seq(100, 1000, by = 100), nboot = 20)
plot(psrange.test)
```

The following simulates covariates with perfect overlap (i.e. the differences between covariate values between treatment and control is random).

```
df.overlap <- getSimulatedData(ncontrol = 1000, nvars=1,
                               treat.mean=0.5, treat.sd=0.4,
                               control.mean=0.5, control.sd=0.4)
psrange.overlap <- psrange(df.overlap, df.overlap$treat, treat ~ .,
```

```
                                samples = seq(100, 1000, by = 100), nboot = 20)
plot(psrange.overlap)
```

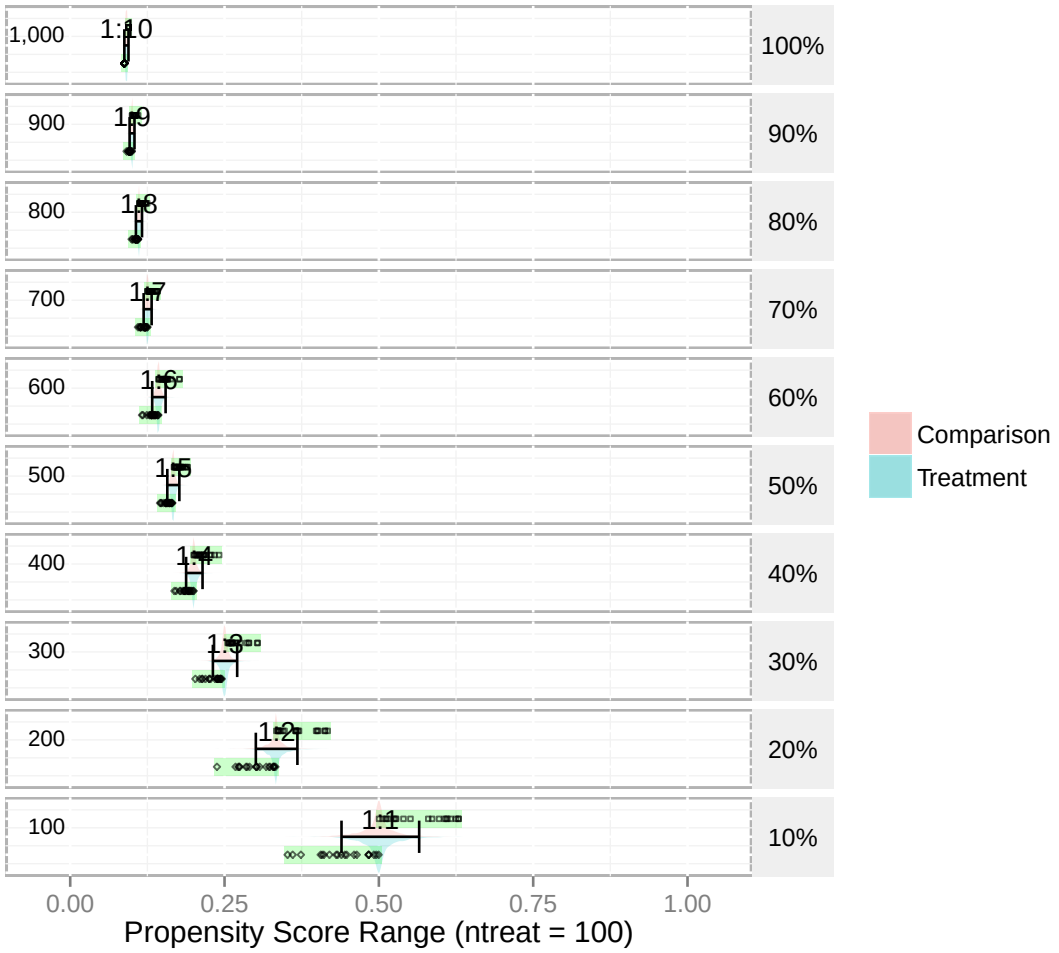


Figure 81: Propensity Score Ranges for Varying Treatment-to-Control Ratios with Perfect Overlapping Covariate

The following simulates covariates with almost no overlap (i.e. the covariate values can almost perfectly predict treatment placement).

```
df.nooverlap <- getSimulatedData(ncontrol = 1000, nvars=1,
                                 treat.mean=0.2, treat.sd=0.4,
                                 control.mean=0.8, control.sd=0.4)
psrange.nooverlap <- psrange(df.nooverlap, df.nooverlap$treat, treat ~ .,
                             samples = seq(100, 1000, by = 100), nboot = 20)
plot(psrange.nooverlap)
```

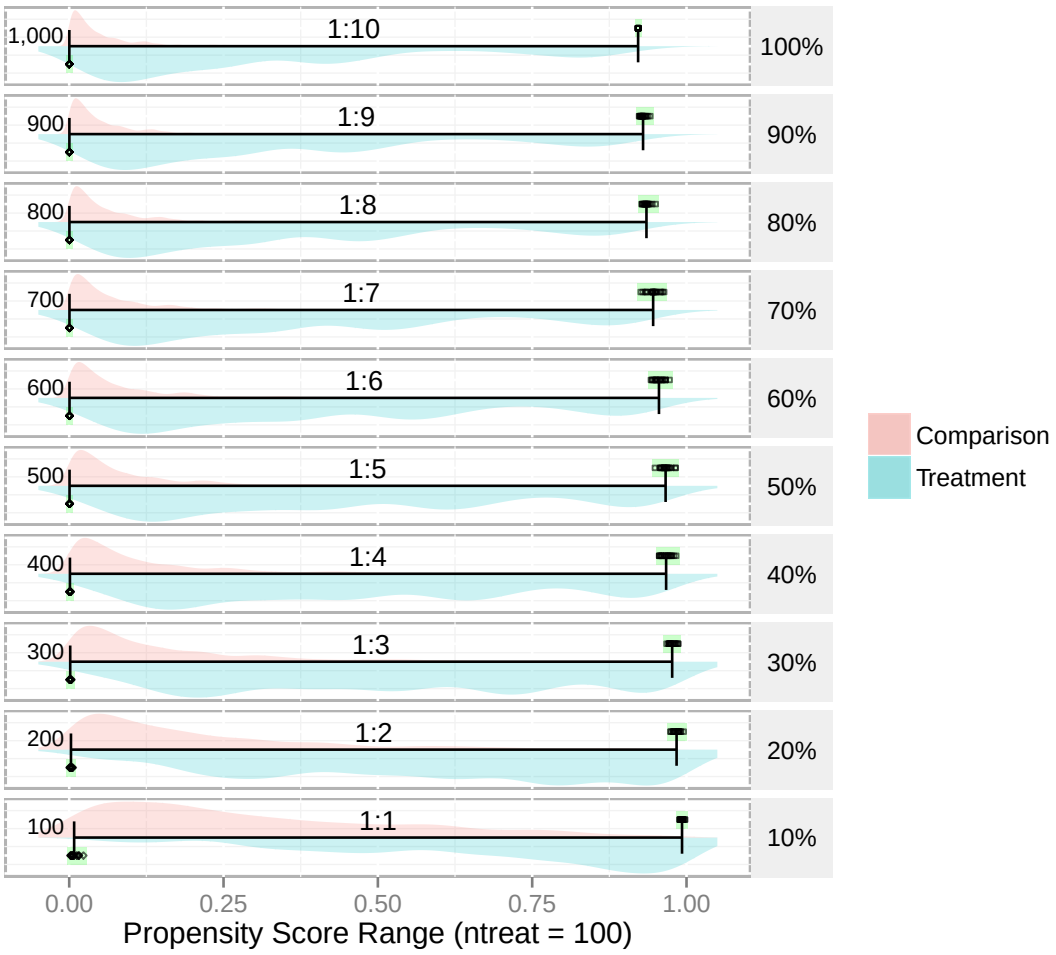


Figure 82: Propensity Score Ranges for Varying Treatment-to-Control Ratios with Non-Overlapping Covariate