November 2017

Investigating the Impact of Student Opt Out on Value-Added Measures of Teacher Quality

Joshua Marland

Investigating the Impact of Student Opt Out on Value-Added Measures of Teacher Quality

A Dissertation Presented

by

JOSHUA J. MARLAND

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2017

College of Education

Educational Policy, Research, and Administration

Research, Educational Measurement, and Psychometrics

Investigating the Impact of Student Opt Out on Value-Added Measures of Teacher Quality

A Dissertation Presented

By

JOSHUA J. MARLAND

Approved as to style and content by:

_____
Stephen G. Sireci, Chairperson

_____
Lisa Keller, Member

_____
Aline Sayer, Member

_____
Cynthia Gerstl-Pepin, Dean
College of Education

ACKNOWLEDGEMENTS

Like all accomplishments in life, this dissertation and all of the work leading up to it were the result of a tremendous amount of support and encouragement from many people in my life. I feel incredibly grateful that I had the opportunity to learn from the people I did throughout my graduate career. What I am most thankful for is that each of my committee members came to each meeting ready to learn more about my topic, but to also teach me through the process.

Aline Sayer gave me the foundation for this topic, and her depth of knowledge and ability to explain complex topics to me has been incredibly helpful throughout the process. I wish her well in retirement!

Lisa Keller has been both a formal advisor in this work, as well as a life advisor to me. She pushed me in what were my most difficult classes, and in any conversation we had, and for that I am forever grateful. I am also thankful for her support in navigating life, and her ability to explain complex technical questions via text message.

Finally, Steve Sireci has demonstrated to me the value of being a lifelong learner. He comes to every conversation ready to learn, and has pushed me to slow down and think through issues in a much deeper way than I would have otherwise. He has also been my biggest advocate throughout my career at UMass, and I hope that I can be the same to others throughout my life.

In addition to my committee, I am incredibly thankful to the entire REMP family. Faculty have always had open doors to me, and are so giving of their time and knowledge. REMP students have really been my everyday support system, from our

lunches together to bowling to digging in on technical homework. I have learned valuable lessons from each and every student in the program.

Andrew Rice has supported me since before I went to UMass, both technically and professionally. He has always provided valuable insight to our conversations, and I look forward to many more. Kristen Huff and Jason Schweid have a lot to do with why I went to UMass, and both have been fantastic mentors to me even when I wasn't aware of the lesson at the time.

My entire family has been involved and encouraging throughout this entire process. My parents always believed I could do anything I wanted, even when I thought I couldn't. They listened to presentations when they had no idea what I was talking about (and still do), and asked questions about what were probably incredibly boring aspects of graduate school, so that I felt supported. My sister, Marybeth, and nephew, Jesse, have been an inspiration to me every day. Marybeth works harder every single day to provide for Jesse than I ever will as a psychometrician, and she is a constant reminder of what is important in life. My grandfather, Freddy, has also been hugely encouraging and a force for me in finishing so quickly – and for that I thank him.

Lastly, I have to thank my partner, Matty, for his constant support and encouragement for the past 8 years! He has never doubted me in any situation, brought me back from the brink on countless occasions, forced me to think about my assumptions (life assumptions - not statistical), and reminded me to relax constantly. He's also been incredibly understanding when I have had to focus on school at all hours of the day and times of the year. I hope that someday I can be as supportive to him as he has been to me for so long.

ABSTRACT

INVESTIGATING THE IMPACT OF STUDENT OPT OUT ON VALUE-ADDED MEASURES OF TEACHER QUALITY

SEPTEMBER 2017

JOSHUA J. MARLAND, B.S., UNIVERSITY OF FLORIDA

A.M., BROWN UNIVERSITY

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Stephen G. Sireci

Student assessment nonparticipation (or opt out) has increased substantially in K-12 schools in states across the country. This increase in opt out has the potential to impact achievement and growth measures used for educator and institutional accountability. This simulation study investigates the extent to which value-added measures of teacher quality are impacted as a result of varying degrees of opt out, as well as various types of nonrandom opt out. Results show that the magnitude of opt out has a greater impact on stability of value-added estimates than the type of nonrandom opt out patterns simulated in this study, with root mean square differences in value-added estimates and standard errors increasing as the magnitude increased. In addition, classification agreement decreased as magnitude increased. Finally, one type of opt out, where the highest achieving students in the highest achieving classrooms did not participate, appeared to have more of an impact on stability than the other types of opt out in this study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

Students choosing to not participate in annual summative assessments (hereafter "opting out") is a relatively new phenomenon in United States K-12 education, with substantial increases in some states and districts over the past several years. Student participation in assessments has important implications: the scores are used for a host of instructional and accountability purposes, including student grade promotion or graduation, and for accountability for teachers, schools, districts and states.

Much of the literature related to opting out of assessments is based on recent news reports or press releases about which students were expected to participate in assessments, and who ultimately did. Reports about the characteristics of students who opt out differ across localities, with some states or districts finding that wealthy, higher-achieving students opt out, such as in Oregon, while New York education officials reported that lower-achieving students in relatively wealthy districts were slightly more likely to opt out. There has also been news coverage of the reasons for which people support the right to opt out. Some reports show that opt out activists oppose the more difficult consortia assessments, or that they do not support evaluating teachers or schools based on the results (Pizmony-Levy & Green Saraisky (2016). Opting out of a state assessment can have implications for many within the education community who rely on assessment results for decision-making purposes.

1.1.1 Federal Participation Requirements

Under Section 1111(2)(I)(ii) of the 2001 No Child Left Behind Act, the United States Department of Education (USDE) required that 95 percent of eligible students participate in the grades 3-8 assessments for English Language Arts (ELA) and Math at the aggregate and subgroup levels. This means that 95 percent of all eligible students were required to participate, as well as 95 percent within each of the federally-protected subgroups, such as English language learners or students with disabilities. Overall and subgroup participation rates were calculated at the school, district, and state levels. Not included in participation rate calculations were the approximately one percent of students with the most significant cognitive disabilities, as long as these students took the state's alternate assessment. As noted in policy guidance from the USDE in 2013, these participation requirements have historically been enforced by the USDE, with states providing regular retake opportunities for absent students (USDE, 2013). The Institute of Education Sciences (IES) found in 2007 that less than one percent of schools did not make their accountability targets because of the participation rate requirement.

State education agencies missing their 95 percent participation rate requirement can face sanctions such as a formal request to comply, a cease-and-desist order, or the withholding or suspending of Title I funds that are meant to support low-income students (Camera, 2015). States are not necessarily required to meet a 95 percent participation rate if they do not receive Title I dollars, which is a major policy lever for USDE. Under the new Every Student Succeeds Act (ESSA) passed in 2015, states are now required to factor low-participation rates into school-level accountability ratings, and to have some level of discretion over how they do so (Ujifusa, 2015). Perhaps in preparation for this

change, the USDE sent letters to 12 states to ensure that they had a plan to address low-participation rates in the assessment at the state, district, or subgroup levels (Klein, 2015).

According to unverified media reports by the National Center for Fair and Open Testing (hereafter FairTest) (2015), at least 14 states had non-negligible numbers of students who chose to opt-out in 2014-15, ranging from approximately 4,600 in Pennsylvania to 240,000 in New York, the latter being the focus of this study. FairTest, along with other organizations, had been advocating for students to opt-out of the newly-implemented national consortium assessments because some believe they are perceived as more difficult than previous assessments, thus leaving students and teachers uncertain about how they will perform (Clark, 2015).

1.1.2 Understanding Federal Testing Requirements

One clear effect of opting out is on participation rates in the state assessment. As mentioned, the participation rate requirements for states set by the USDE in 2001 under the reauthorization of NCLB is 95 percent of all eligible students, including for each federally-protected subgroup (No Child Left Behind Act, 2002). Less clear is the impact that opt out can have on achievement status and growth or value-added measures that states calculate using student assessment scores. The primary goal of NCLB was to have all students, regardless of background, reach the proficient level as defined by each state, by the year 2014 (No Child Left Behind Act, 2002).

As part of their NCLB requirements, states calculated the percent of students in each performance level, as well as the change in percent of students who attained the proficient level across years, for accountability purposes. The percent of students reaching proficiency in the "All Students" group, as well as for each subgroup, in ELA

and Math was considered the status measure, while the change in percent proficient in the same subjects between two years was used as a progress measure (termed Adequate Yearly Progress (AYP)) under NCLB. Through the combination of these two measures, schools were held responsible for improving student performance until all students reached the proficient performance level on the state assessment (No Child Left Behind Act, 2002). As mentioned, missing the participation requirement can mean that a school, district or state fails to meet AYP overall, even if they meet their proficiency targets.

As part of the 2010 Race to the Top (RttT) competition, and the Elementary and Secondary Education Act (ESEA) waivers in 2012, the federal government expanded its focus from school and district accountability to include teacher accountability by requiring that states incorporate measures of student learning into teacher evaluation systems. Most states that participated in RttT or that subsequently received an ESEA waiver chose growth or value-added measures that purported to represent the extent to which students in a classroom grew in an academic year. In reality, these measures are the result of conditional status change calculations, which represent the extent to which a student and/or classroom changed in the distribution of similar students or teachers (Castellano & Ho, 2013). This change was then attributed to teachers for evaluation purposes. Additional information about these methodologies is provided in Section 2.6.

Using these conditional status change measures, rather than achievement status measures, to evaluate teachers was intended to address concerns that teachers of low-achieving students could never be considered effective because their students would have difficulty reaching proficiency in a given year. Theoretically, because these growth and value-added models compare students to other similar students, and/or similar classrooms

to each other, teachers with high concentrations of lower-achieving students could perform well relative to other teachers of lower-achieving students, even though their students did not reach proficiency. However, Newton et al (2010) found that, even after controlling for student-level characteristics, teachers with high concentrations of high-need students tended to receive lower value-added scores with correlations of approximately -0.2 to -0.5. On the flip side, higher concentrations of Asian students and higher levels of parental education were positively correlated with higher value-added estimates. In New York, however, the correlation between demographic characteristics and value-added estimates has been close to zero, perhaps because they included student-level, as well as classroom-level averages for the same characteristics, as covariates in their model (NYSED, 2015).

The comparative nature of these value-added models creates a situation in which student opt out can influence both the accuracy and stability of teacher evaluation measures, depending on several factors, such as the magnitude of opt out and whether opt out patterns are considered to be nonrandom. Random opt out in large numbers could affect the standard errors of any measure created with student assessments because fewer students are likely included in the calculations. Value-added measures typically only include the students in a teacher's classroom for that year, which is roughly 30 or fewer students in a given elementary classroom. To contextualize this number, most states do not calculate results for subgroups with fewer than 30 students because of the instability of the measures with so few observations. Given USDE's stance on the stability of measures with fewer than 30 students in a group, it is safe to say that the number of

students included in value-added estimates is already relatively low for stability standards.

Nonrandomness driven by student-level characteristics may affect the accuracy of teachers with large concentrations of a characteristic. For instance, teacher value-added measures could be biased up or down if all English language learners chose to opt out of the assessment, depending on how systematically different their performance is as a group from other students taking the assessment. Nonrandomness driven by classroom-, school-, or district-level characteristics could impact measures created for each level as well. For instance, value-added estimates could be biased if all higher-achieving students concentrated in certain classrooms opt out of the assessment, because this essentially removes the upper end of the distribution of test takers. This would not only affect these teachers, but would also likely impact teachers with no students choosing to opt-out, because of the comparative nature of the value-added methodology.

This study investigates the impact of nonrandom opt out on student achievement-based value-added measures used for evaluating teachers. Using simulated data generated from real results from one state, I vary the magnitude of opt out in teachers' classrooms, as well as in the overall sample, to determine the impact on reliability and stability of teacher evaluation measures. I then vary the nonrandomness of opt out by relating it to prior achievement in classrooms to determine the impact on teacher evaluation measures for those with and without opt-out students. Finally, I also vary the interaction between magnitude of opt out and nonrandomness to fully investigate the issue.

First, I provide more background on opt out trends in states around the country, then provide a missing data framework in which to situate student opt out, as both are important for understanding how patterns could affect value-added methodologies.

1.1.3 Opt Out in the United States

Rowland-Woods, Wixom, and Aragon (2015) reported that states have in some ways begun to address opt out, either through official statements from state chief school officers advocating that students not opt out, or through tougher laws making it more difficult for students to do so. Some states, like Texas, have an existing law that explicitly does not allow students to opt out of assessments. As described below, however, some state legislatures have advocated for students' rights to opt out, such as in Oregon, Delaware, and New Jersey.

Oregon Department of Education officials reported that approximately five percent of students opted-out in 2014-15, most of whom were non-disabled white students who traditionally perform well on the assessment (Hammond, 2015). After the 2015 assessment administration ended, Governor Kate Brown urged districts to work with parents to stress the importance of assessments and the potential implications of low-participation rates, while at the same time she signed a bill requiring districts to notify parents twice a year of their right to not participate in the state assessment (Ujifusa, 2015). This bill also created two school ratings systems, one of which penalizes schools for low-participation rates, while the other does not.

In Delaware, 10 percent of high school juniors did not participate in the assessment in the 2014-15 school year statewide (Albright, 2015). A bill designed to allow students to opt-out of the state assessment was vetoed by the governor, even after

gaining support from the Delaware Teachers' Union and the State House of Representatives. In New Jersey, where the percent of students opting-out was reportedly just under 10 percent, a bill was introduced in the state legislature that would allow parents to provide written notice to the school that their child would not be sitting for the assessment (Walker, 2015). The bill, however, was not considered when the senate acted on other legislation related to the state's participation in Partnership for Assessment of Readiness for College and Career (hereafter PARCC) assessments (Clark, 2015).

In 2013, as opt out momentum grew across the state, New York State Education Department (NYSED) officials issued guidance to superintendents and principals of all public schools stating that there was no statute or regulation specifically related to allow students to opt-out of the assessment (Katz, 2013). In the guidance, NYSED officials stated that taking state assessments is considered part of the "course of study," and that opting out could negatively impact their child's school or district accountability standing.

According to media reports cited by NYSED, the percentage of students statewide choosing to opt out from the New York state assessment was at its highest level ever in 2014-15, at approximately 20 percent, with estimates as high as 90 percent in some districts on Long Island and in the eastern part of the state (NYSED, 2015). This represented approximately 240,000 fewer students taking the Grade 3-8 assessments in ELA and Math. According to NYSED, opt-out students statewide were more likely to come from average or low-need districts, and were more likely to receive scores in the lowest two achievement levels in ELA or Math (NYSED, 2015). Rice, Marland and Meyer (2016) found that lower-achieving students, based on prior achievement scores, in higher-achieving districts were more likely to opt out in their analysis of 28 districts in

New York. In addition, there was variance across districts in the types of students who were more likely to opt out. Higher-achieving students were more likely to opt out in some districts, while lower-achieving ones did in others. NYSED, like many state education agencies, use assessment scores for a host of accountability purposes, including status and growth measures. As mentioned, both measures can be affected by large proportions of student opt out, which we discuss next.

1.1.4 Missing Data Framework

A large proportion of students opting out represents itself as a missing data challenge in statistical calculations, like the ones performed for creating value-added estimates of teacher quality. In the statistical literature, Rubin (1987) provides a framework for how to evaluate both the pattern and mechanism for missingness. The pattern refers to which data are missing from the analysis, such as randomly across the set of variables used in analysis, which is often how missing data are regarded in the calculation of value-added estimates. Students with missing data might be dichotomously coded as 1 (missing) or 0 (not missing), or students with a current assessment score are dropped from the analysis, so that complete case analysis can be performed.

Because of the large proportion of students opting out in 2014-15, we must now consider other patterns to better understand the extent to which missing data impact value-added estimates. A common missing data pattern from Rubin (1987) is the univariate pattern, where respondents are missing data for only one variable. We might consider this to be the case in 2014-15, where, for the most part, students are only missing the most recent assessment score used for value-added estimation. A second missing data pattern would be the multivariate two pattern, where data are missing for the

same respondents on subsequent measures. This could be the case in 2015-16 if the exact same students opt out of the assessment this year as last year. The last common data pattern is when we have fewer respondents at each subsequent measurement opportunity, such as if opt out continues to grow with fewer students taking the assessment every year.

The other consideration with missing data is the mechanism by which the pattern was created, which is often referred to as one of three categories: missing completely at random, missing at random, and missing not at random (Schafer & Graham, 2002; Rubin, 1987). Each pattern represents a potentially different relationship between the distribution of observable characteristics and missing data, and the reasons for which the data are missing. Each mechanism can differentially affect statistical indices often used for evaluating the quality of a measure – in particular, the reliability and precision of the measure, which in this case are value-added estimates of teacher quality (Schafer & Graham, 2002). Figures 1-3 below help to organize each of the missing data mechanisms into a framework, while focusing on the univariate pattern where we have only one missing data element.

1.1.4.1 Missing completely at random (MCAR)

This can be interpreted to mean that opt out patterns are totally random on observable characteristics, are unrelated to the outcome of interest and with the reason for being missing. This might be represented by similar proportions of students of all demographic groups and/or prior performance categories deciding not to participate in the assessment in 2014-15.

In Figure 1, I begin laying out a framework for organizing the relationship among student characteristic and achievement with their reasons for opting out of the assessment

and whether they do, in fact, opt out. In the figure, we see that student characteristics, such as prior achievement or demographics, are related to student achievement in the current year. We also see that the reason for opting out of the state assessment is related to opting out. However, there is no relationship between student characteristics or current achievement and whether a student chooses to opt out.

1.1.4.2 Missing at random (MAR)

Missing at random can be interpreted as data that may be missing on observable characteristics, such as demographics, and may be related to opt out, but a student's reason for opting out is unrelated to the missing data (student achievement in the current year). For instance, we might see that English Learners (ELs) are less likely to participate in the assessment, but their current achievement (the outcome measure) spans all performance level categories.

In Figure 2, we see that the two previous relationships persist (student characteristics with current achievement and the reason for opting out with doing so), however we now relate student characteristics with opt out. This is meant to represent the fact that characteristics may be related to opting out, but there still does not appear to be a relationship between the missing data and whether students opt out.

1.1.4.3 Missing not at random (MNAR)

Missing not at random builds on the previous two patterns, where the relationship between observables and opt out persists, but there is now a relationship between current student achievement (missing data) and opt out. In this case, we know some characteristics that are related to the reasons student do not to participate in the assessment, and they do, in fact, not participate. For instance, as NYSED reports,

wealthy, white students who were lower achieving in 2014 were more likely to opt-out of the assessment in 2015. In Figure 3, we now relate current student achievement to student opt out, which makes the relationship nonrandom. I should note here that I am not relating student characteristics to the reasons for student opt out because we cannot argue that because a student is white or high-achieving that they chose to opt out. We can only state that:

1. Student characteristics are related to student achievement.
2. Student reasons for opting out are related to whether they do opt out
3. Missingness in current student achievement is related to whether students choose to opt out.

1.2 Statement of the Problem

Student assessment opt out poses several challenges to the educational community, who are reliant on test scores for a host of purposes. The most obvious challenge is that teachers, principals, and policymakers are unable to make inferences about the status of achievement for students who choose to opt out. This is the primary purpose of summative annual assessments. A secondary, more recently-developed purpose of annual assessments is to use the results for evaluating teachers as one measure in multiple measure systems. While some disagree that evaluating teachers using student scores should be done at all, the practice persists across the nation, and likely will continue into the foreseeable future.

To date, very few studies have considered the extent to which opt out patterns and mechanisms can impact value-added estimates. Researchers have typically considered all missing data to be MCAR or MAR, thus ignoring the issue. The magnitude of opt out in

New York and other states necessitates thoughtful investigation into exactly how much, and which, statistical indices are impacted by nonrandom missingness of student scale scores.

1.3 Significance of the Problem

As mentioned, teachers are being evaluated using multiple measures, one of which is value-added or growth estimates. Many states use the results of evaluations for high-stakes decisions, such as hiring, firing, and promotion. To make sound decisions, value-added estimates should, at a minimum, be reliable and precise measures of what they purport to measure. A large degree of nonrandom opt out poses a threat to both indices, thereby threatening the utility of the measures in evaluation systems. In this study, I consider various scenarios for opt out that I believe to represent realistic opt out patterns in several states.

1.4 Purpose of Study

The purpose of the current study is to consider the extent to which teacher value-added estimates are impacted by the magnitude of opt out patterns, as well as by the relationship between opt out and prior achievement.

1. What is the impact of opt out on value-added measures?
    a. How does opt out in different magnitudes within a teacher's classroom impact value-added measures?
    b. How does varying degrees of relationship between opt out patterns and prior achievement impact value-added measures?
2. What is the impact of opt out on classification of teachers value-added estimates using realistic classification systems?

CHAPTER 2

LITERATURE REVIEW

2.1 No Child Left Behind Act of 2001

In 2001, Congress passed the No Child Left Behind (NCLB) Act, which was a reauthorization of the Elementary and Secondary Education Act (ESEA) of 1965. The reauthorization was a departure from existing state assessment and reporting practices in several ways, the first of which was that it included a requirement that schools make academic progress with all groups of students in all schools, including schools not receiving Title I funds for low-income students. The primary goal of NCLB was to have all students, regardless of background, reach the proficient level as defined by each state by the year 2014 (No Child Left Behind Act, 2002).

As part of their NCLB requirements for accountability purposes, states calculated the percent of students in each performance level, as well as the change in percent of students who attained the proficient level across years. The percent of students reaching proficiency in the "All Students" group, as well as for each subgroup, in ELA and Math was considered the status measure, while the change in percent proficient in the same subjects between two years was used as a progress measure (termed Adequate Yearly Progress (AYP)) under NCLB. Through the combination of these two measures, schools were held responsible for improving student performance until all students reached the proficient performance level on the state assessment (NCLB, 2002).

States were also required to report results for each of nine subgroups, including students who were economically disadvantaged, part of a major racial or ethnic group, or English learners, or who had a disability (NCLB, 2002). Under the Improving All

Schools Act (IASA) of 1994, states were only required to make and report progress for all students in schools receiving Title I funding from the U.S. Department of Education (Redfield & Sheinker, 2004). Reporting progress for all students effectively meant that states did not have to disaggregate results for reporting on subgroups of students.

Another new requirement under NCLB was annual testing of at least 95 percent of eligible students in all grades 3-8, and at least once in grades 10-12, in English Language Arts and Math (Redfield & Sheinker, 2004). Under IASA, states were required to assess students annually in one grade in each of three grade spans (Grades 3 – 5; Grades 6 – 8; Grades 10 – 12), but the act did not include a participation rate requirement like in NCLB.

2.1.1 Identification for Improvement

Schools, districts, and states could be identified for improvement if they failed to make adequate progress with any subgroup of students or the "All Students" group, for those not yet at the proficient level on the state assessment. These groups could also be identified for improvement if the participation rate for any subgroup or for the "All Students" group was below the 95 percent requirement, which made ensuring participation in the assessment as important as ensuring students were able to meet grade-level standards. However, most schools found it much more difficult to meet their AYP proficiency targets than the participation requirement (Institute of Education Sciences (IES), 2007).

According to the IES report, 75 percent of schools met their AYP targets for the 2004-05 school year (IES, 2007). For the 25 percent of schools that did not make AYP targets, 43 percent missed for the "All Students" group, 19 percent missed for two or

more subgroups of students, 21 percent missed because of one subgroup, 3 percent missed because of the participation rate requirement, and 14 percent missed for some other reason. This 3 percent of schools that failed to meet because of participation rates translates to less than 1 percent of all schools, which means that schools have historically not had an issue with meeting this particular requirement.

2.1.2 Assessment Exemptions

Under both IASA and NCLB, states were encouraged to utilize universal design principles in developing their assessments so that as many students as possible could participate with or without testing accommodations (Redfield & Scheinker, 2004). While full inclusion was the aim, students with the most significant learning differences where accommodations still did not meet their needs were allowed to not participate in the general assessment. In these cases, states were allowed to administer an alternate assessment of alternate achievement standards to as many students as they deemed necessary, but were subject to what was called the "1 percent rule". Elledge et al (2009) wrote that the "1 percent rule" allows for 1 percent of students in a state or district who score at proficient or above on the alternate assessment to be counted as proficient in the district's or state's AYP calculations. One percent can be 1 percent of all students or 10 percent of students with special needs. USDE also issued interim policy options that would allow up to 2 percent of all students to be counted toward AYP targets, as long as students met proficient or advanced standards on an assessment of modified grade-level achievement standards. This could be 20 percent of students with disabilities or 2 percent of all students.

The National Center on Education Outcomes (NCEO) found in their 2005 analysis of 21 states that the percent of special education students participating in the general assessment ranged from 78 percent in Connecticut to 100 percent in New Hampshire. Cameto et al (2009) report that the proportion of students participating in the alternate assessment has been as high as 15 percent in some states, for example. Elledge et al (2009) also wrote that in 2004-05 almost all states met the 95 participation rate requirement for students with disabilities, with 45 states meeting the goal in reading and 46 in Math. In addition, more than 80 percent of states reported that more than 90 percent of students with disabilities were participating in the general assessment with accommodations when necessary.

2.1.3 Policy Shifts Around the Use of Assessments

As mentioned, NCLB assessments were primarily used for school, district and state accountability purposes through most of the early 2000s. Each state also had its own achievement standards, on which students were assessed annually. Toward the end of the 2000s, however, several notable policy changes took place that expanded the uses of assessments, while also making the assessments of grade-level standards more difficult for everyone.

In 2009 President Obama announced the Race to the Top (RttT) program, which expanded the federal government's focus from school and district accountability to include teacher accountability. USDE stipulated that for states to receive a significant monetary award, they must incorporate measures of student learning into teacher evaluation systems and adopt rigorous academic content standards. Over three phrases of awards, 19 states were given RttT awards, meaning they all committed to implementing

these two new requirements (USDE, 2016a). Shortly after the RttT awards were given in 2010, the USDE also allowed states to apply for waivers from certain ESEA provisions, including the requirement for meeting the 100 percent proficiency target by 2014 that was set under NCLB. Not surprisingly, to receive a waiver, states were subject to the same two requirements from RttT: incorporating student learning measures into educator evaluations and adopting more rigorous academic content standards (USDE, 2016b). Over the next several years, 42 states received waivers, which means that most of the country was now implementing some form of teacher evaluation and the adoption of more rigorous content standards.

Both reforms became relatively controversial within the larger policy context because they were expected to be implemented on a relatively short timeline, while so little was known about either. The statistical models that were used for evaluating educators were in their infancy and not well understood by many, especially by the educators who were being evaluated using them. The new academic content standards that states adopted were made public around the same time as RttT, meaning that teachers needed to learn how to teach their students the content of these new standards, while simultaneously being evaluated on the student assessment results on which the standards were being assessed (Bakeman, 2015; Fairbanks, 2015).

According to the Education Next Poll on School Reform conducted by Henderson, Peter and West in 2016, many supported the Common Core State Standards (CCSS) early in their implementation; however, this confluence of challenges negatively affected support over time. In 2013, 76 percent of teachers supported the standards; this number dropped to 40 percent in 2015. When omitting references to accountability (i.e.

teacher evaluation), teachers were slightly more in favor of the use of the standards in their schools. Members of the public, teachers and parents all reported that the use of the standards in their district had a negative impact. Of those who knew the standards were in use (34 percent of survey respondents), 51 percent reported they had a negative impact. Seventy-three percent of teachers reported that the standards were being used, and 49 percent of them reported a negative effect (Henderson, Peter and West, 2016).

In relation to support for annual assessments on the standards, the same poll found that 67 percent of the public supported annual testing, while 21 percent opposed it. Parents were more supportive than teachers of continuing to assess annually, with 66 percent of parents responding they supported it, while just 47 percent of teachers supported it. With respect to the utility of the resulting scores, only 16 percent of parents reported that a standardized assessment is an accurate indicator of what his or her child learned in the classroom, while 22 percent felt teacher grades were accurate, 25 percent reported that written observations were accurate, and 37 percent reported that actual student work was an accurate indicator of what their child knew.

Taken together, this research shows that about half of parents and teachers are supportive of annual testing, with a small proportion reporting that standardized assessment scores are an accurate indicator of what a child knows. In addition, roughly half of parents and teachers supported the CCSS, which were now being assessed on annual tests. This lack of support for each of the reform initiatives is, perhaps, what led some to gravitate toward the opt out movement.

2.2 Who supports the opt-out movement?

As students are the ones ultimately being assessed, it is they who must opt out of a state assessment. However, the media reports and research about the reasons for opt out have focused primarily on the actions of policymakers that may have led people to question state assessments and their associated uses, and on teachers and parents for influencing children to opt out. To date, only a handful of media reports exist around the reasons for which students chose to opt out, and focus mostly at the high school level. While opting out at the high school level has taken place in great numbers, it also occurred in substantial numbers in lower grades in several states. Given this phenomenon, some researchers have investigated the adults who support the opt out movement, under the hypothesis that they may in turn influence the magnitude of opt out in their localities.

Pizmony-Levy & Green Saraisky (2016) conducted a survey of approximately 1,600 opt out activists in early 2016 as a means to better understand parent motivations for supporting the opt out movement and for allowing their children to opt out. The authors found that respondents to their survey who were considered opt out activists, because they frequented opt out websites, tended to be predominantly white women, and relatively more educated and wealthier than the U.S. general public. Average income of respondents was $125,000, compared with a median of $53,000 for U.S. households. In addition, 97 percent reported having completed postsecondary education, with almost 60 percent reported as having a graduate degree. Finally, 45 percent of respondents reported they were teachers or educators, and another 16 percent reported having teachers or educators in their circle of friends.

Most activists reported positively about their own schools, either their child's or the ones where they work. Sixty-eight percent responded they would give their own school an A or a B, which is more positive than the U.S. general public, where 51 percent gave schools in their community the same grades. The authors hypothesized that this could be due to one of two situations (or both): Activist respondents are wealthier, and have access to what most would consider better quality schools in their neighborhoods, and/or they reject the current popular notion that schools in the U.S. are failing.

According to Pizmony-Levy & Green Saraisky (2016), 44 percent of educators, which made up half of the respondents in their survey, reported they did not support the use of test scores in teacher evaluation. Thirty-two percent reported that standardized tests force teachers to teach to the test, 22 percent reported that standardized tests take away valuable instructional time, and 18 percent did not support the implementation of CCSS.

Henderson, Peter, and West (2016) reported in their poll that 32 percent of parents supported the right to opt out of a state assessment and 52 percent opposed this right. Teachers felt similarly to parents, with 32 percent supporting the movement, and 57 opposing it. Members of the public (respondents without children) were less likely to be supportive of allowing students to opt-out, with 25 percent supporting the idea and 59 percent opposed to it.

In a poll conducted by Phi Delta Kappa and Gallup in 2015, 41 percent of parents supported the right to opt out of a state assessment. When asked about their own children, only 31 percent of parents reported they would opt their child out of the state assessment, and 59 percent reported that would not opt their child out. Differences in favorability

toward opt-out existed across demographic groups, with Black and Hispanic respondents less likely to be in favor of it when compared with White respondents. Bennett (2016) and others argue that this may be due to differences in educational quality across demographic groups, where a standardized assessment score serves as an objective measure that can shine a light on disparities.

Pizmony-Levy & Green Saraisky (2016) asked activists whether they opted their own children out of a state assessment, and found that 63.3 percent reported opting all of their children out, and 11.2 opted some of their children out. Most of the parents also reported they would likely opt their children out in the future as well (82.8 percent very likely; 9.3 percent likely).

2.3 Federal Policies on Opting Out

The future of opting children out depends, in large part, on the implications for doing so. The only federal requirement is that 95 percent of eligible students participate in the state assessments, with exceptions for students with disabilities who should participate in an alternate assessment. For states that do not meet the participation requirement, the federal government can impose one or several sanctions, depending on the magnitude of the issue. In 2015, USDE officials wrote to 13 state education agencies, asking how their states planned to handle low participation, and outlined the potential sanctions that exist if the issue persists (Ujifusa, 2015). In their letter, officials offered potential ways the state might address low participation, including:

1. "Lowering a local education agency's (LEA's) or school's rating in the State's accountability system or amending the system to flag an LEA or school with a low participation rate;

2. Counting non-participants as non-proficient in accountability determinations;
3. If the State has received ESEA flexibility, identifying a school that misses participation rate targets over multiple years as a priority or focus school;
4. Requiring an LEA or school to develop an improvement plan, or take corrective actions to ensure that all students participate in the Statewide assessments in the future, and providing the SEA's plan to review and monitor such plans;
5. Requiring an LEA or school to implement additional interventions aligned with the reason for inadequate student participation, even if the State's accountability system does not officially designate schools for such interventions;
6. Designating an LEA or school as "high risk," or a comparable status under the State's laws and regulations, with clear explanations for the implications of such a designation; and/or
7. Withholding or directing use of State aid and/or funding flexibility."

State education agencies could choose one of the above options for addressing low participation in schools to ensure that the issue does not persist (Chism, 2015). In the event that issues do persist, the USDE also outlined in their letter to states the potential sanctions against the education agency that exist for low participation:

- A formal request that a state comply;
- Increased department monitoring of a state;
- Conditions on federal title I aid provided for low-income students;
- Placing a state on "high-risk" status;
- Issuing a cease-and-desist order;
- Entering into a compliance agreement with a state;

- Withholding all or a portion of a state's Title I administrative funding; or
- Suspending, and then withholding, all or a portion of a state's Title I grant.

In her letters, USDE Assistant Secretary Deborah Delisle wrote that states missing the assessment participation threshold could also face a loss of funding for other programs, including monies from the Individuals with Disabilities Education Act, School Improvement Grants, programs for English language learners, rural schools and migrant students, as well as Title II, which funds professional development and training for teachers (Camera, 2015). Because of the increase in student opt outs and the increased federal pressure to address the issue, state policies and laws related to opt out have begun to change.

2.4 State-Level Opt Out Policies and Activities

Lorenzo (2016) reported that state opt out policies generally fall into four groups:

1. Opt out is prohibited (34 states and Washington D.C.);
2. Opt out is permitted completely (California and Colorado);
3. Refusal is permitted or opt out is permitted with constraints (10 states); and
4. Opt out policy is left to local districts (Idaho, Montana, Nevada, and South Dakota).

Rowland-Woods, Wixom, & Aragon (2015) reported specific state-level policies on student and parent rights for not participating in the state standardized assessment. The authors found that opting out is currently technically allowed in statute in Utah and California, and may have been in law in New Jersey and North Dakota. However, New Jersey's current policy prohibits schools and districts from opting students out of the state

assessment – and the bill that was introduced did not make it passed the state legislature. There are other states where the state department of education allows students to opt-out, such as in Minnesota or Michigan (though Michigan advises against it). When Rowland-Woods et al published their paper, students in Oregon could exercise a religious exemption, which is one of several available exemptions to students in different states. Other exemptions include physical disability, medical reasons, or emergencies. In Texas and Arkansas, participation in the state assessment is mandatory. For the most part, guidance to parents typically cites section 111 of NCLB, which states students should take part in the state assessment.

According to Bennett's 2016 research across states, the greatest proportion of opt-out took place in New York, where it was about 20 percent in ELA and Math. Table 1 contains rates for the states listed in the Bennett (2016) report. Rhode Island, Colorado, and Maine all had rates higher than the 5 percent rule set by USED as part of NCLB. Bennett reports that the rate of refusal in high schools was also much higher than at the elementary grades. In Washington state the 11th grade refusal rate was 49 percent in ELA and 53 percent in Math, where over all grades, the rate was 2 and 3 percent, respectively. He also reports that the high school refusal rate was the primary reason states were put on alert by USED for low-participation rates.

FairTest (2015) provided the number of students that opted out across states as well, the estimates of which were based on media and direct reports from people working in a particular state. Overall, FairTest estimated that more than 670,000 students opted out of state assessments across the country, though that number has not been independently verified for all states on their list. They report that 240,000 students opted

out in New York, 130,000 in New Jersey, and 100,000 in Colorado. Eight additional states had more than 10,000 students opt out in 2014-15. More detailed examples from several states follow.

2.4.1 Washington

Administrators from the Washington State Board of Education published a presentation of assessment results for 2014-15, where they cite three primary drivers for opt out on state assessments: a new assessment system, new learning standards, and organized opposition to testing in general, and the Smarter Balanced Assessment Consortium (SBAC) and Partnership for Assessment of Readiness for College and Career (PARCC) assessments in particular (Parr & Teed, 2016). In their presentation, Parr & Teed write that approximately 22 districts had participation rates lower than 30 percent on the 11th grade Math Smarter Balanced Assessment. They also find that there was a negative relationship between the percent of test refusals and the percent of free or reduced lunch meals in a district.

In addition, they found that about 20 percent of the 2,162 schools in Washington with data did not meet the federal participation rate requirements. Most of those that failed were high schools, where 95 percent failed to meet the requirement. The median participation rate for elementary and middle schools was 95 percent, where it was around 15 percent across subjects at the high school level.

2.4.2 New Jersey

Harris & Fessenden (2015) report that in New Jersey, about 4.6 percent of students in grades 3-8 did not participate, while almost 15 percent did not in Grade 11. Perhaps because of the magnitude of opt out, New Jersey legislature introduced a bill that

laid out procedures for students to opt out of the state assessment, so as to avoid confusion during the PARCC testing window.

In the bill, schools would have been required to provide students with an alternate activity during testing while others are taking the test. The bill also would have required that school districts provide information to parents about the subject area, the administration dates, the manner in which the results were going to be used, and whether the assessment is required by the federal or local government. The legislature ultimately passed a nonbinding resolution requiring the state education commissioner to provide guidelines to school districts for students who refuse to take the assessment.

Officials from the New Jersey Department of Education released an "action plan development guide" in late 2015, meant to provide districts with specific steps for improving participation for schools or subgroups with low participation in the 2014-15 school year (New Jersey Department of Education, 2015). The guide advises districts to perform a root cause analysis of the reasons students may have chosen to opt out, by attempting to answer such questions as:

- "How many students did not take the assessments because they refused or their parents would not permit them to take the test? What were their concerns?"
- "Did we have a large number of voids? If so, what were the reasons (e.g., students discontinued taking the test, students did not respond to a sufficient number of questions to get a valid score)?"

The guide also provides strategies that may help to address some of the reasons students chose to opt out. For instance, one strategy in the guide was to have administrators meet with a PARCC special education advisory group to discuss available

accommodations for special needs students. Another strategy was to have staff meet directly with parents to support the interpretation of PARCC score reports.

2.4.3 Connecticut

In a presentation released by the Connecticut Department of Education (CDE), officials argue that assessment participation is a matter of educational equity, and that "inferences derived from assessment results is partially dependent on the percentage of students who participated in the assessment" (Butler & Gopalakrishnan, 2016). In the presentation, administrators point to federal and state law that requires at least 95 percent participation in the assessment, though their stated goal is 100 percent. They also specify four categories for district or school participation rates that could lead to the withholding of funds. Similar to New Jersey officials, Connecticut outlines the reasons for which students may not participate in the assessment, and offer strategies for addressing those reasons.

2.4.4 New York

Prior to the increase in opt out, and in anticipation of it, NYSED released a policy memo to district superintendents and principals of all public schools, summarizing state and federal policy on student participation in the state assessment (Katz, 2013). In the memo, a state assessment administrator wrote that the state assessment is considered to be part of the course of study for students. He also stated there is no provision that allows for students to opt out of the state assessment, and doing so may have negative consequences for schools and districts. Finally, he mentioned that participation in field tests is also important for the reliability and validity of the operational tests.

Two years after the release of the memo, NYSED (2015) reported that approximately 240,000 fewer students took the assessment in 2014-15. They also reported that those who chose to opt out were in primarily white middle- and upper-income districts, centered in and around Long Island. In New York City, the refusal rate was just 2 percent. In addition, students who refused were more likely to be lower achieving in the prior year when compared with those who participated in the assessment.

Chingos (2015) used publicly available data to determine the statewide opt-out rate in New York, which he reported to be 28 percent across the 648 districts in his dataset, and 21 percent when he weighted the calculation by district enrollment. He reported that this downward trend implies that larger districts tended to have lower opt-out rates. He also reported that 20 percent of districts had opt out rates lower than 10 percent, while 13 percent had a majority of their students (> 50 percent) opt out. Harris (2015) corroborates these findings in reporting that in 60 districts, the number of refusers outnumbered the number of test-takers.

In his analysis, Chingos also found that the percent of students receiving free/reduced lunch (FRL), student enrollment, and student prior achievement all decreased as opt out rates increased. To take the analysis further, he regressed the percent of opt-out in a classroom on percent FRL and average test scores and found that the positive correlation between test scores and opt-out became negative when FRL is taken into account. His interpretation was that a one standard deviation increase in test scores was associated with a 7-percentage point decline in the opt-out rate, when controlling for FRL.

Rice, Marland, and Meyer (2016) found in their analysis that schools with higher opt out rates likely experienced an increase in achievement scores due to nonparticipation of lower-achieving students. They estimated the difference between opt out and non-opt out students to be approximately .2 standard deviation units on the score scale in Math and .12 in ELA. In addition, they found that there was a small but significant difference in student growth percentiles from the prior year between those who opted out and those who participated in the assessment. Across grades and subjects, the authors report a 2-point difference in student growth percentiles between the two groups.

American Institutes for Research (AIR) conducted an in-depth analysis of opt out for NYSED as part of their value-added contract with the state to ensure that teacher value-added estimates were unaffected (AIR, 2016). They found that opt out was non-random, with lower proportions of English Learners and economically disadvantaged students in classrooms with higher proportions of opt out. They reported there was a strong relationship between average participation and prior achievement across grades (with the exception of Grade 6). Almost 50 percent of teachers had non-participation rates lower than 10 percent (Table 2). Classrooms with higher proportions of non-participation tended to have higher prior achievement when compared with classrooms with zero or low non-participation. Similarly, high non-participation classrooms had lower proportions of high-need children.

This finding by AIR may seem in contradiction to what NYSED reported in their earlier analysis. At the student level, NYSED found that students who opted out were more likely to be at the bottom two levels on the state assessment, whereas AIR found that classrooms with high proportions of opt out tended to have higher average prior

achievement. These two facts taken together can be interpreted to mean that students who opted out tended to be lower achieving students in high-achieving classrooms or districts. In their analysis, Rice, Marland, and Meyer (2016) analyzed opt out trends at the student, school and district levels, and arrived at a similar finding: lower-achieving students in higher-achieving classrooms or districts chose to opt out of state assessments in New York.

Finally, teachers with high proportions of non-participation were also less likely to be rated as effective or highly effective in the previous year. An in-depth explanation of New York's analysis of the impact of opt out on value-added measures will be provided in section 2.5.1.

2.5 Missing Data Framework

As mentioned in Section 1.1.2, there are several considerations in determining how best to handle missingness in data sets, considerations such as item and unit nonresponse, univariate and multivariate missing, and the mechanisms by which the data are missing. Rubin (1987) offers a framework for determining the types and magnitude of nonresponse, as well as a variety of methods for addressing nonresponse in surveys, all of which apply to assessments.

2.5.1 Item nonresponse and unit nonresponse

Item nonresponse is when a person responds to most of an assessment, but omits a response to one or several items. Students often fail to respond to items on a state assessment for a variety of reasons, such as running out of time or lack of knowledge. With respect to opt out, there is a possibility that students chose to exercise their right after they began taking the assessment and failed to respond to a string of items. In this

case, students would still likely receive a score, unless the teacher recorded the student's assessment as an anomaly. Unit nonresponse is when a person does not respond at all to an assessment. As mentioned previously, there are a variety of reasons students may not have a valid score for an assessment, but this study primarily focuses on those students who are completely missing because they chose to opt out of the assessment.

2.5.2 Univariate vs. multivariate missing

Also discussed in Section 1.1.2, students missing only the most recent assessment score would fit the univariate missing data pattern, because they are only missing one data element. Students missing several data elements would be considered multivariate missing, which could be the result of students missing consecutive assessment scores, or those missing the current year assessment score and other predictor variables used in calculations. For instance, a student could have all the assessment scores necessary for calculating change over time, but may be missing the demographic controls often found in value-added calculations. This pattern does present itself in value-added estimates currently, and is often modeled through the use of missing flags for each demographic indicator (1 = missing and 0 = not missing). Another possibility that will likely become more prevalent this year as opt out persists in some states is that students may be missing two consecutive assessment scores.

2.5.3 Missing Data Mechanisms

The other consideration with missing data is the mechanism by which the pattern was created, which is often referred to by one of three categories: missing completely at random, missing at random, and missing not at random (Schafer & Graham, 2002; Rubin, 2014). Each pattern represents a potentially different relationship between the distribution

of observable characteristics and missing data, and the reasons for which the data are missing. Each mechanism can differentially affect statistical indices often used for evaluating the quality of a measure – in particular, the reliability and precision of the measure, which in this case are value-added estimates of teacher quality (Schafter & Graham, 2002).

This study is primarily interested in univariate unit nonresponse, which based on analyses performed by NYSED and other states, we can assume to be missing not at random. Rubin (1987) writes that nonresponse can result in less efficient estimates because of the reduction in the number of respondents in a data set. Bias in estimates is also likely because nonrespondents are different from respondents, and it is difficult to know the exact reasons why nonresponse exists.

The definition of nonresponse expands from when respondents choose not to provide information to when editing procedures reduce the number of responses (unlikely responses to a particular question). The definition is further expanded to situations in which nonresponse is the result of the instrument design – for instance, only those who answer yes to a particular question are exposed to parts of the instrument.

2.5.1 Addressing Nonresponse in Data

There are two common practices for addressing missing data: discarding cases where values are missing, or imputing values based on known information for respondents. Nonresponse for a small proportion of respondents is fairly common, and could potentially be discarded from analysis. For instance, in past years in New York, a small proportion of students did not have a current assessment score and were not included in value-added estimates for teachers. There is potential that exclusion could

reduce efficiency and increase bias, but a small proportion is likely to only have a small impact on overall results.

Rubin advocates for imputation, either single or multiple imputation, which could be utilized in nonresponse situations. Both forms of imputation utilize known information for respondents, or for the sample, to estimate missing values for respondents. Single imputation is the estimation of one response for respondents with missing data. For instance, this might be used to estimate a current or prior assessment score for students who opted out based on their assessment scores. The key here is that only one assessment score is estimated for that student. This is an attractive technique because it allows for relatively straightforward complete data set analysis. Utilizing single imputation is slightly problematic, however, in that the score is the result of an estimation, and analysis does not take the additional variability due to nonresponse into account. To address this, Rubin advocates for the use of multiple imputation.

Multiple imputation is similar in that the technique estimates missing data elements for respondents, but is different in that it imputes multiple possible values for each missing data element. The technique is meant to address sampling variability that would exist in single imputation. For instance, we might estimate five current year assessment scores for students who opted out this year. Analysis becomes less straightforward, however, because it requires acknowledging the sampling variability that exists from the imputation process.

However, to date, NYSED and other states have chosen to treat students who opt out as missing from the data, meaning they calculate value-added for teachers only with those students who took the assessment in the current year. According to AIR (2016),

Rubin's methods require that strong assumptions be made about the reasons for nonparticipation. AIR also argues that utilizing a multiple imputation framework ignores differential instructional effects for teachers, i.e. two teachers with similar groups of students might have very different instructional effects on their students, which would not be picked up when imputing assessment scores. For these reasons, NYSED completely omitted scores for opt out students, and performed a comparison of model fit with and without opt out students, as well as a comparison of mean growth percentiles (MGPs) and classifications for teachers in the previous year with and without students who opted out in the current year.

AIR (2016) computed the r-square for the 2013-14 model with (complete model) and without the non-participation students from 2014-15 (incomplete model), and found very similar results between them. The r-square was approximately .7 across grades, with differences no larger than .01 between the complete and incomplete models. At the student level, the root mean square of the difference between the two model predictions was never larger than .5, which the authors argue translates to one half of one scale score point. The correlation of growth percentiles at the student level for those with student growth percentiles was .999 in the complete and incomplete models.

Teacher MGPs in 2013-14 calculated with and without non-participating students in 2014-15 were correlated at about .98, suggesting a strong linear relationship between them. The relationship between the change in a teacher's MGP and classroom characteristics appeared to not be large and/or systematic for most characteristics. Teachers with large positive changes in MGP tended to have lower proportions of

economically disadvantaged students. Similarly, large positive changes in MGPs were also related to lower non-participation rates.

AIR (2016) also calculated the mean standard error for the complete and incomplete models, and found that the incomplete model had slightly less precision. The mean SE/SD for the complete model were 4.18 and 10.92, respectively, while for the incomplete model they were 4.57 and 11.16. The mean SE/SD for the complete model was .38, and .41 for the incomplete model. Overall, the authors report that a teacher could expect to have a difference in their MGP of 2 points.

According to AIR, 82 percent of teachers were expected to get the same classification rating used by the state under both models as a result of opt out, 3.7 percent were expected to increase one rating category, 4.3 percent were expected to decrease one rating category, and about .1 percent were expected to move both up and down by two rating categories. Almost 3 percent of teachers would have expected transitions from the top two categories in the complete model to the bottom two in the incomplete model (without opt out students).

Classification agreement between the complete and incomplete model effectiveness ratings across all teachers was high – at 93 percent. This is higher than the expected rate of classification agreement AIR reported in the previous section. Teachers in categories of consequence, however, had lower agreement – only 80 percent of teachers in the bottom two rating categories in the complete model remained in the same category in the incomplete model. Similarly, 90 percent in the top two categories in the complete model did not change categories in the incomplete model.

2.6 Value-Added Estimates of Teacher Quality

As outlined in Section 1.1.2, growth has typically been defined as an increase in the percentage of students reaching the proficient mark on the state assessment. Because this was seen as a crude growth metric, policymakers began searching for a method that gave schools and teachers credit for improving student learning while also holding them accountable for students reaching proficiency.

A variety of growth methodologies exist and are employed for school and educator accountability purposes, with calculations and interpretation ranging from incredibly simple to complex. Relatively simple student growth scores can be calculated on assessments employing a vertical scale, where a common scale is utilized across grades (Castellano & Ho, 2013). The common scale is linked to a developmental continuum for mastery of a single domain, such that scores in one grade can be subtracted from a subsequent one to represent growth across years for students. However, most state summative assessments do not employ a vertical scale, possibly because of operational challenges associated with maintaining the developmental continuum across years, thus precluding them from utilizing this gain score model for holding teachers and schools accountable.

Because of the lack of a vertical scale, growth model adoptions by states over the past 10 years or so belong to a family of models where changes in performance from one year to the next are calculated for each student and compared to demographically or academically similar students and are termed "conditional status models." (Blank, 2010; Castellano & Ho, 2013). Value-added models (VAMs), or "residual gain models", are

one of the members of this family, along with student growth percentiles and multivariate models. VAMs use linear regression to estimate expected changes in current test scores, given past student scores, with the result represented as a residual gain (or loss) on the assessment scale. Many VAMs, like New York's, include classroom or school characteristics as well, which changes interpretation of the difference between predicted and actual scores to include similar classrooms or schools as well.

Meyer (1995) argued that school achievement measures were flawed and provide an inaccurate picture of changes in performance over time. In his paper, Meyer posited that VAMs provide valuable information about the extent to which schools could improve student outcomes, after student, classroom, and school characteristics are controlled for in a multi-stage regression model. The central argument to the method is that once factors outside of the classroom are accounted for in the model, the residual difference between actual and observed achievement can be attributed to a stakeholder in the educational system.

2.6.1 Specifying a Value-Added Model

There are two common implementations of value-added regression models for teacher evaluation, and choosing one or the other depends on assumptions made by the developer. The first model is parameterized as follows:

$$A_{ig} = \lambda_1 A_{i,g-1} + \lambda_2 A_{i,g-2} + \lambda_3 A_{i,g-3} + E_{ig}\beta + e_{ig} \qquad (1)$$

Where A_{ig} is the current year test score for student i, $\lambda_1, \lambda_2, \lambda_3$ are slope parameters, $A_{i,g-1}$, $A_{i,g-2}$, $A_{i,g-3}$ are prior year scale scores, E_{ig} are indicator variables for specific

teachers, β are teacher estimates of effectiveness, and e_{ig} is the student-level error term (Guarino, Reckase, Stacy and Wooldridge, 2014). This model is often used because it is relatively straightforward for calculating teacher effects, and allows for the calculation of teacher-level standard errors. In addition, the use of a teacher fixed effect, β, controls for nonrandom assignment of students to teachers. Conceptually, the teacher indicators partial out the effect of the teacher from other covariates in the model, which some argue are related to assignment of students to teachers (Clotfelter, Ladd, & Vigdor, 2007; Dieterle et al, 2012; Kalogrides, Loeb, & Beteille, 2013). Variations of this model are used in practice in several places, including New York City and Hillsborough County, Florida.

Another common VAM implementation is similar to Equation 1, except teacher fixed effects, β, are not included in the model. Instead, teacher random effects are used in a multi-level model, where students are considered to be nested within a classroom. Teacher value-added estimates are constructed by averaging the student-level residuals within a classroom. Because this model does not include teacher fixed effects, it assumes random assignment of students to teachers. New York and Florida both use a form of this model, designed by AIR, for evaluating teachers, principals, and schools. Papay (2011) and Newton et. al (2010) found that school-level value-added estimates derived using these two implementations had Spearman rank correlations between .88 and .92, which can be interpreted to mean they achieve similar results, but at least some teachers will be ranked differently depending on the choice of model.

In both value-added model specifications, states and districts can (and do) include covariates considered important for empirical and policy reasons (McCaffrey et al, 2004).

The use of student, teacher, or school-level covariates changes the inferences one can make from the results of the models because similar groups are compared with each other, thus requiring interpretation to be contextualized as relative to similar peers. As mentioned, this is an attractive requirement to policymakers who are responsible for evaluating teachers, because teachers of high-achieving students are compared to each other, as are teachers of lower-achieving students. In general, researchers have found that the inclusion of covariates beyond prior achievement has had little effect on teacher effectiveness estimates (Ballou et al, 2004; Papay, 2011). In New York's random effects model, student-level indicators were included for whether the student lived in poverty, or received special education or English language supports (AIR, 2015). In addition, classroom averages of each demographic characteristic were also included.

2.6.2 Stability of Value-Added

In a white paper, Kane and Staiger (2010) investigated the stability of teacher value-added estimates in New York City and Los Angeles schools. The authors found that correlations of value-added estimates tended to be low to modest across years, ranging from .35 to .5 in different studies they performed. Newton et al. (2010) found similar correlations in ELA across grades, and slightly higher correlations in Math at .43 to .63 across grades.

To investigate the practical implications of differences in value-added estimates across years, Kane & Staiger created quartile categories as a classification scheme, and found that teachers at the bottom and top quartiles generally tended to stay there. In New York City, 68 percent of top quartile Math teachers remained in the top quartile in consecutive years, and none moved to the bottom quartile. Similarly, no bottom quartile

teachers moved to the top quartile in the next year. When averaging value-added estimates for teachers across two years, 80 percent who were in the top quartile in Math remained there, and none moved to the bottom quartile. The classifications were less stable in the middle two quartiles, where 44 percent of teachers in the third quartile stayed there in a subsequent year, and roughly one quarter moved up or down a quartile. Estimates become more stable with more data in the middle quartiles. Kane and Staiger found that 56 percent of teachers who start in the third quartile remain there in the next year, and 20 percent move up or down a quartile.

In Los Angeles, teachers in the top quartile in Math in their first year moved students up .14 standard deviations in their 3rd and 4th year. This can be interpreted to mean that the highest value-added teachers in year one were averaging positive value-added estimates in the third and fourth years of the study. Bottom quartile teachers saw similarly-sized drops in achievement in their students in their 3rd and 4th year. Similar to New York City, the predictive power of value-added estimates improved with more data. Teachers in the top quartile in their first two years moved students up .17 standard deviations above similar students. Teachers in the bottom quartile after two years lost ground with their students, who averaged a loss of .18 standard deviations below similar students.

McCaffrey et al. (2009) found that approximately 25-35 percent of teachers in the bottom quintile of value-added estimates remained there in a subsequent year, while 10 – 20 percent moved all the way up to the top quintile, depending on the grade. Similarly, of those who started in the top quintile, between 25 and 35 percent stayed there while approximately 10 – 15 percent moved to the bottom quintile. Intertemporal correlations

of teacher value-added estimates across years at the elementary and middle levels were about .2 to .3 in the cities in Florida that were studied.

In their 2010 study, Newton, et al. classified teachers by deciles, and investigated the extent to which there were changes across model specifications, courses, and years. Models were specified with and without fixed effects, with and without demographic characteristics as covariates, and as a three-level model with demographics. Across model specifications, 56 – 80 percent of teachers changed at least 1 decile, between 12 and 33 percent change by 2 or more deciles, and 0 – 14 percent changed by 3 or more deciles. Across years, 74 to 93 percent of teachers changed by 1 or more deciles, 45 to 63 changed by 2 or more, and 19 – 41 changed by 3 or more deciles.

2.6.3 Missing Data in Value-Added Estimates

One study, by Papay in 2011, investigated whether missing students could impact a teacher's value-added estimate. The author calculated value-added for teachers using two different assessments that were administered in the spring in the same years, meaning this is a Spring to Spring calculation for the state assessment and the Scholastic Reading Inventory (SRI). He correlated the value-added results derived from both assessments for teachers and found there to be a .44 correlation, which is a low to modest correlation. Arguing that students could be missing for any number of reasons (such as mobility or absence), Papay then restricted his sample to only those students who had data for all administrations, and found the correlations improved to .54 for the Spring to Spring models using the state assessment and SRI assessment.

McCaffrey, et al. (2011) utilized data from an urban school district, and found that only 20 percent of the 10,332 students who attended school there in a five-year period

had complete data for all years. They then studied two possible ways that missing data could impact value-added estimates: missing teacher links to students, and missing student scores over time, the latter of which is more relevant to the current study. Missing teacher links can be problematic in longitudinal situations where one is attempting to model the persistence of teacher effects on students over time, in which case a valid teacher-student link for each year is required.

To investigate the impact of missing teacher links, the authors employ three methods for generating value-added estimates (termed teacher effects in McCaffrey, et al.). The first is to set missing teacher effects to zero, which means that missing teachers had no effect on student learning in future years. The second method samples missing teacher effects from the same distribution as observed teachers. The third draws missing teacher effects from their own distributions, with variance components for the distributions estimated separately from the observed teachers. The authors calculated three sets of current teacher effects, and found .99 correlations for each grade and subject. In addition, the authors found a similar mean and variance for the three sets of teacher effects. This aspect of missing data only applies in situations where historic teacher effects are modeled as a function of current achievement for students, essentially serving as one covariate in the value-added model. In practice, this model is not typically used because of the data requirements McCaffrey, et al. mention above.

The authors also simulated missing scores to investigate the missing at random assumption often employed in calculating value-added estimates. To do so, McCaffrey, et al. simulated teacher effects for 250 teachers, with 50 per grade, which were centered to have a mean equal to zero and variance to 13 percent of the residual variance for each

year. They used teacher effects to generate 100 samples of five years of test score data for 1,250 students in classes of 25 students to which they were randomly assigned. They then use a probability model to predict the number of scores students that would be missing from a value-added model. In their simulation, students with lower prior achievement scores were expected to have more missing scores, which is typical in achievement data. The authors calculated complete value-added and with missing data, and found that the estimated teacher effects were relatively robust to missing data. Correlations between the true value-added and the average teacher effect across all 100 data sets was higher than .99. Teachers with very small and very large true effects tended to have the largest deviation from the average value-added estimates. McCaffrey, et al. write that this is likely due to the Bayesian shrinkage used in estimation, which essentially weights teacher effects by their reliability.

Karl, et al (2013) utilized a correlated random effects model to determine the extent to which missingness impacted teacher effects in a K-12 setting as well as in a university. At the K-12 level, the authors modeled approximately 6 percent missing data for 304 fourth, fifth and sixth grade teachers, and calculated teacher random effects. The authors found that the relationship between random effects from the MAR condition were correlated above .99 to the MNAR conditions. In addition, they found that random effects did not predict whether students attended on the day of the test. However, the authors argue that while value-added estimates were robust to missingness in their study (and in McCaffrey, et al. (2011), missingness should not be ignored and practical implications explored. While correlations were high, large differences could exist for even a small number of teachers as a result of missing data.

As can be seen, very few studies have explored the extent to which missing data impacts teacher value-added estimates. Interestingly, the general working assumption by policymakers is that there is very little missing data; however, both Papay and McCaffrey found that only 20 to 50 percent of students have complete data for all years prior to the current year. None of the existing literature addressed a phenomenon like opt out, where a large magnitude of students were missing in current data, and could be missing in future year calculations without a tenable solution for modeling missing priors. My study aims to add to the literature around the impact of missing current achievement scores in value-added estimates.

All of the studies published to date have not explored a phenomenon such as opt out, where the magnitude was large and the reason for opting out unknown. The current study aims to contribute to the literature in that respect.

CHAPTER 3

METHODS

3.1 Methods Overview

A simulation study was conducted to examine the amount of bias introduced into value-added estimates under various opt out conditions, and to determine the extent to which opt out impacted classification of teacher effectiveness measures. Observed scale scores were simulated to represent students' test scores on a typical statewide assessment for four grades, hereafter referred to as grades 3, 4, 5, and 6. The probability of opting out was simulated using parameter estimates from empirical data. Students were then identified for de-selection from the analysis randomly, based on the probability of opting out of the assessment, and based on their prior achievement. Grade 6 observed value-added estimates were calculated in separate models using grades 3, 4, and 5 as conditioning years.

The data were generated using a multivariate sampling approach from Castellano and Ho (2014) to produce a nested structure observed in real data; that is, students nested within classrooms. Furthermore, to simulate realistic data, the parameters in this simulation were based on real test data from empirical analysis from New York. Correlations, standard errors, and root mean square differences were investigated across conditions to better understand the extent to which error is a function of opt out. Quartiles of value-added estimates were used to determine rates of agreement between complete and incomplete value-added estimates across replications for every classroom.

3.2 Data Generation

3.2.1 Generating observed scale scores

Scale scores were generated using a multivariate normal sampling approach utilized in Castellano & Ho (2014) and with parameter estimates from empirical data, where within- and between-classroom deviations were sampled from multivariate distributions and summed to create student-level observed scores. This sampling allowed observed scale scores to be generated for each student in a classroom with the addition of a common classroom effect. The multivariate sampling procedure begins with Equation (2):

$$\begin{pmatrix} Y_g^B \\ X_{1g}^B \\ X_{2g}^B \\ X_{3g}^B \\ O_g^B \end{pmatrix} \equiv \begin{pmatrix} Y_g^B \\ \mathbf{X_g^B} \\ O_g^B \end{pmatrix} \; N_5(\boldsymbol{\mu}, \boldsymbol{\Sigma^B}) \qquad (2)$$

where $\boldsymbol{\mu}$ is a vector of average scale scores across all students in the generated data for each year, as well as the percent of students who will opt out of the assessment. The dimensions of the average scale score and percent of opt out matrix $\boldsymbol{\mu}$ are 5 x 1. $\boldsymbol{\Sigma^B}$ is the variance-covariance matrix for the average classroom scale scores and opt out with dimensions of 5 x 5. Y_g^B is the current year classroom deviation from the average score, $\boldsymbol{X_g^B}$ is a classroom deviation for each of the three prior years, and O_g^B is the classroom deviation for opting out. For scale scores, $\boldsymbol{\mu}$ was set to 310 for each year because average

scale scores in a state tend to remain relatively stable across years, assuming tests are not vertically scaled and have within-grade scales. For the probability of opting out, this was varied across three conditions: 5, 10, and 20 percent. The covariances between opting out and scale scores was set to zero for all years, except for the immediate prior year.

In Equation (3), we have the multivariate sampling procedure for within-class deviations.

$$\begin{pmatrix} Y_{ig}^{W} \\ X_{1ig}^{W} \\ X_{2ig}^{W} \\ X_{3ig}^{W} \\ O_{g}^{W} \end{pmatrix} \equiv \begin{pmatrix} Y_{ig}^{W} \\ \mathbf{X_{ig}^{W}} \\ O_{g}^{W} \end{pmatrix} \; N_5(\mathbf{0}, \boldsymbol{\Sigma^{W}}) \tag{3}$$

Where **0** is a 5 x 1 column vector that represents the average of the within-classroom deviation for each of the four generated years of scale scores and probability of opting out, and $\boldsymbol{\Sigma^{W}}$ is the 5 x 5 variance-covariance matrix for generating student-level deviations from the classroom mean for each year. Y_{ig}^{W} is the within-classroom deviation for the current year, and $\boldsymbol{X_{ig}^{W}}$ represents the deviation for each of the prior years, and O_{ig}^{W} is the within-classroom deviation for opting out. To generate current and prior year scale scores and the probability of opting out, I used Equations (4), (5) and (6):

$$Y_{ig} = Y_{g}^{B} + Y_{ig}^{W} \tag{4}$$

$$\boldsymbol{X_{ig}} = \boldsymbol{X_{g}^{B}} + \boldsymbol{X_{ig}^{W}} \tag{5}$$

$$O_{ig} = O_{g}^{B} + O_{ig}^{W} \tag{6}$$

Where Y_{ig} is the current year test score for each student in a classroom that is the sum of the between and within-group deviations, Y_g^B and Y_{ig}^W, and $\boldsymbol{X_{ig}}$ is a matrix containing each of the three prior scale scores, the dimensions of which are $N \times J$, where N equals the number of students and J equals the number of prior scores. O_{ig} is the probability that a student will opt out of the state assessment in the current year.

Data generation required the use of student- and teacher-level correlations of scale scores across years, the student-level standard deviation of scale scores, and the intraclass correlation observed in real data, which is the proportion of the variance attributed to classroom-level differences in scale scores. Intertemporal correlations of scale scores at the student level were set to 0.85 between adjacent years, 0.83 for scores with a two-year lag (i.e., current with two years prior, one year prior with three years prior), and 0.75 for scores with a three-year lag (i.e., current with three years prior.)

For generating the probability that a student opts out, I used the student-level correlation between the dichotomous indicator for opting out from the empirical data and the immediate prior year scale score, the correlation between average prior achievement and percent of students opting out in a classroom, the student-level standard deviation of the dichotomous indicator for opting out, and the intraclass correlation of opting out. The dimensions of the correlation matrix that includes scale scores and probability of opting out, denoted as **R**, are 5 x 5. Correlations between all years of scale scores and opting out were set to zero, with the exception for the correlation between opting out and the immediate prior year.

Correlations across years at the teacher-level were set to 0.90 for adjacent years, 0.85 for scores with a two-year lag, and 0.80 with a three-year lag. The correlation

between average prior achievement and the percent of students opting out in a classroom was set to -0.05. These correlations are expressed in a 5 x 5 matrix denoted as $\mathbf{R^B}$. Student-level standard deviations were set to 35 for scale scores, which were held constant across years. The student-level standard deviation of opting out was set to .4. Scale score and opting out standard deviations are expressed in a diagonalized matrix, **D**. The intraclass correlation, expressed in $\boldsymbol{\omega}$, was set to 0.22 and also held constant across years for scale scores, and set to 0.13 for opting out. The $\boldsymbol{\omega}$ matrix is a diagonalized 5 x 5 matrix.

First, the total variance-covariance matrix, Σ, which includes within- and between-classroom differences, was calculated using the student-level standard deviations in **D**, and intertemporal correlations in **R** in Equation (7) as follows:

$$\mathbf{\Sigma = DRD} \tag{7}$$

Next, the between-classrooms variance-covariance matrix, $\mathbf{\Sigma_B}$, were calculated, using the intraclass correlation contained in $\boldsymbol{\omega}$, the student-level standard deviation in **D**, and the classroom-level intertemporal correlations in $\mathbf{R^B}$ in Equation (8):

$$\mathbf{\Sigma^B} = (\sqrt{\boldsymbol{\omega}}\mathbf{D})\mathbf{R^B}\left(\sqrt{\boldsymbol{\omega}}\mathbf{D}\right) \tag{8}$$

The difference between these two matrices results in the within-classroom variance-covariance matrix $\mathbf{\Sigma^W}$ that was used to generate the student-level within-class deviations from the classroom average in Equation (9):

$$\mathbf{\Sigma^W = \Sigma - \Sigma^B} \tag{9}$$

3.2.2 Estimating Value-Added Measures of Teacher Quality

To estimate value-added measures of teacher quality, I used a common method also utilized in Guarino, Reckase, Stacy and Wooldridge (2014) that estimates a teacher effect through the use of dichotomous indicators for each of the 1,000 teachers. The model is parameterized as follows:

$$A_{ig} = \lambda_1 A_{i,g-1} + \lambda_2 A_{i,g-2} + \lambda_3 A_{i,g-3} + E_{ig}\beta + e_{ig} \qquad (10)$$

Where A_{ig} is the current year test score for student i, $\lambda_1, \lambda_2, \lambda_3$ are slope parameters, $A_{i,g-1}$, $A_{i,g-2}$, $A_{i,g-3}$ are prior year scale scores, E_{ig} are indicator variables for specific teachers, β are teacher estimates of effectiveness (fixed effects), and e_{ig} is the student-level error term. This model was used because it is relatively straightforward in calculating teacher fixed effects, and allows for the calculation of teacher-level standard errors. Variations of this model are also used in practice in several places, such as Los Angeles Unified School District and Hillsborough County Schools in Florida.

3.2.3 Empirical Data

Four years of empirical assessment and demographic data for 32,722 students in grades 3 - 8 in 122 schools in 28 districts in New York were made available to generate parameter estimates for this simulation study. The included school years spanned 2011-12 to 2014-15. Because these data represent a subset of the state, I compared several generating parameter estimates with publicly-available data from the state's website. As mentioned in Section 3.1, I intend to generate 6th grade scale scores in Math, so I

restricted my sub-sample data to only those students, which resulted in 8,023 students (Table 3). This is approximately 4 percent of the state's total 6th grade population.

In my sample, average Math achievement for those with scores in 2015 was 308, compared to 304 statewide. While not explicitly modeled in my analysis, I also compared demographic characteristics for students in my sub-sample with the rest of the state. In the sub-sample, 51.9 percent of students are identified as living in poverty, while only 37.6 percent of students are statewide. The percent of English language learners is similar at 7.9 for the sample and 7 statewide. Lastly, the percent of students with disabilities is 15.7 statewide and 13 in the sample data.

In sum, my sample data appears to contain slightly higher-achieving students than the rest of the state, at least for those who participated in the assessment in 2015. Students in my sample also appear to opt out to a slightly lesser extent, which could possibly be due to differences in the way I identified opt outs, which is explained in the next section. Finally, a greater proportion of students in my sample live in poverty than in the rest of the state, and they appear to be slightly less likely to be identified as having a disability than the rest of the state.

3.2.4 Identifying Students as Opt out in Empirical Data

As part of the requirements under the evaluation process in New York State, each district must report to the state teacher-student linkage information that identifies a teacher-of-record for every student. Also included are the number of minutes a student is in a teacher's classroom during the course of the year. The state uses this information when calculating growth measures for use in evaluation, and returns to each district a

data file with a reason for why the student was or was not included in growth calculations.

Using the empirical data provided by the state, students were considered opt outs if they were linked to teachers for the entire year but had no valid current year test score. This means that the student was in a tested grade and linked to a teacher, but had no valid test score for the same year. Students who did not meet minimum enrollment and attendance duration requirements were dropped from analysis, unless they were also identified as not having a valid current year test score, in which case they were also considered opt outs. All students with valid test scores were considered test-takers for the purposes of the analyses.

3.3 Simulation Conditions

As mentioned, the mean probability of opt out was set to 5, 10 and 20 percent, which is meant to simulate varying degrees of opt out in the data. I simulated 100 data sets for each condition for a total of 300 data sets. In each data set, students were chosen for opt out (or de-selected from analysis) randomly (Condition 1 below), based on their probability of opting out generated in the previous steps (Conditions 2), or based on their place in the prior achievement distribution (Conditions 3 & 4). There are four conditions for each magnitude of opt out that simulate possible real-life scenarios across states:

1. Students are excluded randomly across the achievement distribution so that 5, 10 or 20 percent of students are excluded;
2. Students with the highest probability of opt out across the prior achievement distribution were chosen so that 5, 10 or 20 percent of students are excluded;

3. Fifty percent of students chosen for opt out are the lowest achieving students in the top quartile of prior achievement classrooms; the remaining 50 percent are from the other three quartiles of prior achievement; and
4. Fifty percent of students chosen for opt out are the highest achieving students in the top quartile of prior achievement classrooms; the remaining 50 percent are from the other three quartiles of prior achievement

Value-added estimates were calculated once with all students in the analysis for the 100 complete data sets, and once for each of the four conditions with students deselected from analysis. This results in 500 value-added estimates of effectiveness for teachers: one complete value-added estimate, and four based on opt out simulation conditions.

Lastly, the number of students associated with teachers was varied to have a mean of 30 and standard deviation of 10, which mirrors a typical elementary school classroom. The minimum classroom size was 1 student and the maximum was 63. However, as noted later, classrooms with fewer than 11 students were excluded from stability analysis. The total sample size of students per replication was approximately 30,000, and the total sample size of teachers per replication will be 1,000.

3.4 Data Analysis

The relationship between the classroom-level complete and incomplete value-added estimate was examined using the Pearson correlation coefficient. Correlations between complete and incomplete value-added estimates and prior achievement were examined as a check on whether teachers with higher-achieving incoming students could

expect to receive higher VA estimates. I also provide descriptives for the distributions of value-added estimates obtained under each condition; in particular, means, variances and kurtosis. While I anticipate means to remain relatively stable, it is possible that variances may change as a result of missing data, as well as the peak of the distribution.

The stability of the complete and incomplete value-added models was examined by calculating the root mean square difference between the incomplete and complete value-added models across replications from AIR (2016):

$$RMSD = \sqrt{n^{-1}\sum_{i=1}^{n}\left(VA\_Incomplete_{jk} - VA\_Complete_{jk}\right)^2} \qquad (11)$$

The difference between the incomplete and complete VA estimate was calculated for each classroom, *j*, as well as *k* replications ($k = 1, 2, \ldots, K$), where K=100. The result will be on the value-added scale, which is represented in standard deviation units. I also calculated the root mean square difference between value-added estimates obtained under the missing at random condition (#1) and estimates from the MNAR conditions (#2, #3, and #4).

I also examined the extent to which standard errors of teacher fixed effects increase as a result of students opting out of the assessment, by finding the average standard error and the root mean square difference for both the complete and incomplete models.

Lastly, to investigate the practical implications of opt out on value-added estimates, I classified teachers into four rating categories using quartiles, with teachers classified based on their complete and incomplete value-added estimate for each of the data sets. Quartiles were used because they often are in research settings, so as not to

adopt issues inherent in various classification schemes. The number and proportion of classifications in agreement between complete and incomplete value-added estimates across all replications and conditions was calculated, as well as weighted Kappa.

CHAPTER 4

RESULTS

First, I provide descriptive detail on the simulated scale score data that was used to generate value-added estimates, as a means for ensuring the data approximate desired and realistic conditions. Included in this section is the distribution of the percent of opt out in each teacher's classroom across conditions and magnitude. I also provide detail on the average prior achievement for students in each of the four opt out and three magnitude conditions, and I provide correlations for the relationship between prior achievement and opt out conditions at the student and teacher level.

Next, I provide detail on the distribution of value-added estimates across each of the conditions, including means, standard deviations, and a measure of kurtosis, which is also provided as a means for ensuring that the value-added models approximate results from practical settings. Finally, correlations, root mean square difference of value-added estimates and standard errors, and classification agreement are presented as a means for exploring stability under the various opt out and magnitude conditions.

I use the following terminology throughout to refer to the types and percent of opt out: "Opt out conditions" refer to the four types of opt out that are simulated. These are:

- the "random" condition refers to students who were dropped randomly from analysis;
- "highest probability" refers to the condition where students were dropped based on the probability of opting out, which was predicted using prior achievement;

- "lowest achieving" refers to the condition where 50 percent of opt out students had the lowest prior achievement in the top quartile of prior achievement of all classrooms, and the other 50 percent of opt out students were randomly selected.
- "highest achieving" refers to the condition where 50 percent of opt out students had the highest prior achievement in the top quartile of prior achievement of all classrooms, and the other 50 percent of opt out students were randomly selected to opt out; and

Finally, the term "magnitude", when applied to a condition, refers to the 5, 10 and 20 percent opt out conditions that were simulated for each of the four opt out conditions.

4.1 Distribution of Percent Opt Out Across Conditions and Magnitudes

In an effort to make clear the extent to which each opt out and magnitude condition impacts opt out rates, I first provide more information about the distribution of the percent opt out for teachers across conditions and magnitude. I also utilize opt out categories found in the AIR Technical report (2016) so the reader has a comparison point.

The percent of opt out across magnitude and opt out conditions appears to be what I would expect, with the greatest proportion of teachers generally falling into the magnitude category that was being simulated (Tables 4 –7). For instance, in the random 5 percent category, we see that 64.2 percent of teachers have between 0 and 10 percent (Table 4). In the 20 percent random condition, 67.5 percent of teachers have between 10 and 25 percent opt out in their classrooms (Table 4). In both the random and highest probability conditions, the proportion of teachers with more than 50 percent of students opting out is relatively low, with the exception of the 20 percent condition when students are selected based on the probability of opting out (Tables 4 & 5).

In the highest and lowest achieving conditions, the proportion of teachers with more than 50 percent of opt out students is substantially more in the 20 percent condition than in the other opt out and magnitude conditions (Tables 6 & 7). In the highest achieving student condition, 13.8 percent of teachers have more than 50 percent of students opting out, and 14.3 percent have more than 50 percent in the lowest-achieving student condition. This difference is because I purposely selected half of the opt out students (10 percent of the total students) to be from the top quartile of classrooms, while the other half were from the other three quartiles, which concentrates a substantial amount of students in approximately 250 of the 1,000 simulated classrooms, and creates more opportunity for teachers to have high proportions of opting out. Figures 4 –7 display the distributions of the percent of opt out for each condition and magnitude.

In Table 8, I list the number and percent of teachers with exactly zero students in each opt out condition. In this table, we can see that, of the 100,000 teachers (1,000 teachers in 100 replications), there are very few teachers who drop out of analysis fully because all of their students were chosen as opt outs. The highest achieving condition has the greatest number of teachers, with 43 in the 10 percent condition and 37 in the 20 percent condition.

In addition, I also included in Table 9 the number and percent of teachers with 10 or fewer students after de-selection. As mentioned previously, these teachers were excluded from all of the proceeding stability analyses because states typically include all students and teachers with data in estimation, but do not provide effectiveness ratings for teachers that do not meet a threshold for a minimum number of students. In my analysis, I required that teachers have more than 10 students because that is the minimum number

used in some states. Table 9 shows that the highest and lowest achieving conditions had the highest number and proportion of teachers with 10 or fewer students. Nearly 11 percent of teachers in the 20 percent condition were excluded in the highest achieving condition, and just over 9 percent were excluded in the lowest achieving.

4.2 Average Prior Achievement Across Conditions

As a check on the data generation process, I provide more information on the average prior achievement of students across opt out and magnitude conditions. In the random condition, average prior achievement is the same for students who opted out and for those who did not (Table 10). We can see that the average is 310 across all three magnitude conditions as well.

In the highest probability condition, average prior achievement is slightly lower for students who opted out across all three magnitude conditions. The average is 303 for opt out students and 310 for those who did not opt out in the 5 percent magnitude condition. I expect opt out students, in this case, to be lower achieving because I negatively correlated the probability of opting out with achievement in the immediate prior year. However, the average increases slightly for the 10 and 20 percent conditions because I am selecting more students with lower probabilities of opting out in order to achieve the desired magnitude.

In addition, the probability of opting out is much higher for students who did opt out, when compared with those who were not selected to do so in the data generation process. In the 5 percent condition, the average probability of opting out is .85 for opt out students, compared with .15 for non-opt out students. The probability of opting out is slightly lower as the magnitude of opt out increases from 10 to 20 percent. In the 10

percent condition, the average probability is .79, and .74 for 20 percent conditions. Again, to achieve the desired magnitude, it was necessary to select students with lower probabilities for opting out.

In Table 10, we can also see that the average prior achievement for the highest achieving students who opt out is 355 in the 5 percent magnitude condition, compared to 307 for students who did not opt out. We see substantially higher achievement for opt out students in the 10 and 20 percent conditions when compared to the students who were not selected for opting out. Figure 10 provides a visual for the achievement of these students, where the distribution of achievement is almost bifurcated. This bifurcation is the result of selecting approximately half of the opt out students to be the highest achieving in the top quartile of classrooms, while the other half were randomly selected from the other three quartiles of achievement. Because of this targeted selection, we see what approximates a normal distribution for lower-achieving students, and a sharp increase in achievement for opt out students at the higher end of the achievement distribution.

For the teacher level in the highest achieving condition, Figure 14 displays the percent of opt out as a function of average prior achievement for each magnitude. In the figure, we can see that, for the most part, percent opt out is relatively similar across the range of classroom average prior achievement, except that we see a steep increase in opt out beginning around the 315 scale score in each magnitude. This stands to reason, as we purposely selected higher achieving students to opt out, so we would expect higher percentages of opt out at the high end of the achievement scale.

Finally, average prior achievement for opt out students in the lowest-achieving condition is 290 in the 5 percent magnitude condition, compared with 311 for students

who do not opt out. Similar to the highest-achieving student condition, prior achievement increases as the magnitude of opt out increases, with an average of 296 in the 10 percent condition and 306 in the 20 percent condition. In Figure 11, we can see that the prior achievement distribution is also bifurcated, except the increase in percent of opt out is at the lower end of the achievement scale. In addition, Figure 15 displays the teacher-level percent of opt out as a function of average prior achievement for the lowest achieving condition. Here, we see a steep increase in percent opt out at around 325 on the achievement scale, which then tapers off as prior achievement increases.

4.3 Correlations Between Prior Achievement and Opt Out

In Table 11, we see the student- and teacher-level correlations between prior achievement and opting out. At the student level, the correlation is between the dichotomous opt out and each student's prior achievement. At the teacher level, the correlation is between average prior achievement and the percent of students who opted out.

In the random condition, the student- and teacher-level correlations are 0 for the 5, 10 and 20 percent magnitude conditions. In the highest probability condition, student- and teacher-level correlations are slightly negative, which was by design and based on empirical data. In the highest achieving condition, the student-level correlations between prior achievement and opt out range from .31 in the 5 percent condition to .39 in the 20 percent condition. At the teacher level, the correlations range from .64 in the 5 percent condition to .74 in the 20 percent condition.

Finally, the student-level correlations between prior achievement and opt out in the lowest achieving condition range from -0.14 to -0.06 across the three magnitude

conditions, and range from .45 to .6 at the teacher level. The last two conditions (highest and lowest achieving students) were meant to serve as extreme examples where opt out is a function of prior achievement at the student and teacher levels. In addition, the lowest achieving condition was meant to represent an example of Simpson's paradox, where we have a negative correlation at the student level and a positive one at the teacher level, which represents a higher level of aggregation.

4.4 Value-Added Distributions

As mentioned, value-added estimates where teachers had 10 or fewer students are not included in the following analysis. In Table 12, we see that the mean value-added estimate is 0 across all conditions (opt out and magnitude), which is by design. Scale scores were entered into the value-added equation as z-scores, which mean centers them with a standard deviation of 1. The resulting coefficient estimates should also have a mean of zero as well.

The standard deviations shown in Table 12 are similar across opt out and magnitude conditions, and are similar to, but slightly lower than, what we would see in empirical data. In the simulated data, we see standard deviations around .22 for each of the conditions, except in the highest achieving condition where the standard deviations range from .23 in the 5 percent condition to .25 in the 20 percent condition. In empirical data, typical standard deviations range from .25 to .3.

In Table 12, kurtosis remains fairly consistent across all simulation conditions, with values around 3 in each. We can interpret this to mean that the distribution of value-added estimates is similar to that of a normal distribution, at least with respect to the density around the mean estimate.

4.5 Stability of Value-Added Estimates

4.5.1 Correlations

As a first step toward investigating stability of value-added estimates I calculated the Pearson correlations of complete estimates with the incomplete estimates from each of the four opt out conditions and three magnitude conditions for a total of 12 correlation coefficients. In Table 13, we see that the correlations are all higher than .99 for the 5 and 10 percent magnitude condition. The correlations range from .981 - .983 for the 20 percent condition, which is only slightly lower than the other two magnitudes.

4.5.2 Root Mean Square Difference of Value-Added Estimates

As outlined in the methods section, I calculated the root mean square difference (RMSD) between the complete value-added estimates with all students included, and for the incomplete value-added estimates for each opt out condition (4 conditions) in each of the three magnitude conditions. The RMSD's for the random condition serve as a baseline, by which we can compare estimates from the other three conditions to determine the extent to which the simulated nonrandomness impacts the estimates.

As we see in Table 14, RMSD's increases in each condition as the magnitude of opt out increases, with an average of .02 for the 5 percent magnitude condition across replications, .03 for the 10 percent condition, and .04 for the 20 percent condition. An RMSD of .04 in the 20 percent condition represents an average difference in value-added estimates of almost .2 of a standard deviation (Table 14), which is sizeable. This can be interpreted to mean that a teacher in New York, for instance, could expect to move up or down .2 of a standard deviation in value-added estimates if 20 percent of students opt out.

Across opt out conditions, we see that RMSD's are fairly consistent, with the exception of the highest achieving condition, where we see a slight increase in RMSD over the random condition. This increase of approximately .006 - .008 in RMSD across each magnitude of opt out is fairly minimal, but does represent a difference that is due to this type of nonrandomness, where 50 percent of the students opting out are the highest achieving in the highest achieving classrooms.

In an effort to demonstrate how the differences in value-added estimates are a function of prior achievement, I present in Figures 16 – 19 the average absolute difference in value-added between complete and incomplete data sets by prior achievement for each of the four opt out conditions. In the figures, the actual difference in value-added is presented, and not the RMSD's, because RMSD's are calculated for each replication, and prior achievement is fairly consistent across replications by design. I also present differences in value-added as a function of the percent of opt out in Figures 20 –23.

In Figure 16, we can see that the difference in value-added estimates between the complete and incomplete data sets in the random condition is close to zero for much of the prior achievement distribution. At the lower end of the distribution (< 250 on the scale), we see that the difference increases slightly. Similarly, there is a slight increase in the difference in value-added in the 20 percent condition when prior achievement is around 375 on the scale. The plot is similar in the highest probability condition, where the differences in value-added are close to zero throughout the prior achievement scale (Figure 17). In both situations, however, this increase in the difference in value-added may represent a relatively small number of classrooms.

In the highest achieving condition, differences in value-added are slightly negative at the lower end of the prior achievement distribution, and positive at the higher end (Figure 18). This can be interpreted to mean that value-added is slightly lower for the lower achieving classrooms when a substantial proportion of higher achieving students in the top quartile of classrooms opt out. In addition, higher achieving classrooms have slightly higher value-added when other higher achieving students opt out.

In the lowest achieving condition, differences in value-added are fairly consistent across the prior achievement scale, which is similar to the random and highest probability conditions (Figure 19). Differences are also larger at the tail ends of the achievement scale, which is similar to all the other opt out conditions.

4.5.3 Root Mean Square Difference of Standard Errors

Similar to RMSD's of the value-added estimates, I calculated the RMSD of the standard errors (SE's) for teacher fixed effects. We see in Table 14 that the RMSD's of standard errors also increase as the magnitude of opt out increases. In each of the opt out conditions, average RMSD's of the standard errors was approximately .003-.004 in the 5 percent magnitude condition, .005-.006 in the 10 percent condition, and .04 in the 20 percent condition. While the RMSD of SE's was fairly consistent across the types of opt out, it was slightly higher in the highest achieving condition similar to the RMSD's of value-added estimates.

In the random and highest probability conditions, the differences in standard errors are fairly consistent across the prior achievement scale (Figures 24 and 25). However, we see in Figure 26 a substantial increase in standard errors at the high end of the prior achievement scale for the highest achieving condition, which occurs in all three

magnitude conditions as well. The differences are as large as .08 in the 20 percent condition, which represents more than one third of a standard deviation of fixed effects. In the lowest achieving condition, we see a slight increase in standard errors at the same place in the scale where we saw a significant increase in the percent of opt out, which can be interpreted to mean that standard errors increase with large magnitudes of opt out (Figure 27). The increase in standard errors here is smaller than in the highest achieving condition, at around .035 in the 20 percent condition.

Classification Agreement: As a final investigation into the stability of the value-added estimates, I created quartiles of the complete and incomplete fixed effects to approximate "effectiveness" categories of teachers. I then calculated the percentage of teachers where the complete and incomplete quartile rating were in agreement, and averaged that percentage across replications. In Table 14, we see that average classification agreement is similar across each of the opt out conditions, and is higher in the lower magnitude conditions. For instance, in the 5 percent magnitude condition, we see that average classification agreement is 91 percent across all opt out conditions. In the 20 percent condition, average classification agreement was 80.6 percent in the random condition, and drops to 77.7 percent in the highest achieving condition. This can be interpreted to mean that an additional 3 percent of teachers would be misclassified with this type of nonrandomness present when compared to random opt out.

I calculated the average percent of teachers that remained in the same category or moved as many as three quartiles up or down, and present this information by prior achievement quartile (Tables 15 –18). This is meant to better understand the extent to which classification agreement is a function of prior achievement, and to understand just

how extensive the moves are across quartiles under each condition. In Table 15, we see that average classification agreement in the random condition is similar across prior achievement quartiles. In addition, the average number of teachers who move more than one quartile is less than one, even in the 20 percent condition. In the highest probability condition, the average number of teachers who move more than one category is slightly higher – though still less than one across all prior achievement quartiles and magnitude conditions (Table 16). The average number of teachers moving two quartiles is highest in the bottom two quartiles of prior achievement, at approximately .15 (or .1 percent of teachers).

In the highest achieving condition, the average number of teachers moving more than one category is 1.15 – or .7 percent of teachers – in the top quartile of prior achievement (Table 17). The number and percent of teachers moving more than one category in the other prior achievement quartiles is close to zero. Perhaps not surprisingly, it is the top quartile of prior achievement in all three magnitude conditions where we see the highest misclassification when compared with the other three quartiles. For instance, in the 20 percent condition, the average number of teachers moving up or down one category is approximately 24 (or 10 percent of teachers in the quartile) in the bottom three prior achievement quartiles. However, in the top prior achievement quartile, the average number of teachers who would change classification is 44 across replications, or 26.5 percent of teachers in the quartile.

In addition, because the percent of opt out is higher for this quartile, there are fewer teachers for whom a change can be calculated. We can compare this quartile, where approximately 166 of a possible 250 teachers were included in calculations, to the

other three quartiles, where an average of 240 teachers were included. This may also contribute to the substantial difference in classification agreement for this quartile, because teachers whose complete fixed effect was in the top quartile may drop out of the analysis in the incomplete condition, which potentially affects the relative place in the distribution of fixed effects for other teachers.

Lastly, to better understand exactly which teachers changed categories, as well as why they may have changed, there are two additional visualizations for each opt out condition in the 20 percent magnitude condition. Figures 32 –35 show complete and incomplete value-added estimates for each opt out condition and for each teacher by their change in classification quartile. In all of the figures, we can see two trends across all of the opt out conditions. The first is that there is a strong linear relationship between complete and incomplete value-added estimates, which we also could have estimated given the strong correlations. The second takeaway is that teachers who do change quartiles are ones with value-added estimates close to zero. We do not see teachers with high value-added estimates changing classification quartiles to a large extent. While some teachers may have had value-added estimates change substantially when students opt out, they do not necessarily change quartiles.

The final set of figures (Figures 36 –39) are meant to highlight two points, and perhaps makes one point better than the other. The figures show complete value-added and average prior achievement by change in quartile, but also includes teachers who were excluded from stability analysis. The inclusion of excluded teachers in this figure was meant to demonstrate in greater detail just who gets excluded from analysis and its potential impact on classification. In the figures, we can see that teachers who change

quartiles in the random and highest probability conditions are spread across average prior achievement, but have value-added estimates close to zero. However, excluded teachers tend to have more spread in value-added estimates, which may, in part, be due to the fact that there were already a smaller number of students included in those value-added estimates.

For the highest and lowest-achieving conditions, we see a similar result with spread across average prior achievement, as well as value-added estimates close to zero for those who change quartiles. However, we see that excluded teachers tend to be concentrated at the higher levels of average prior achievement for both opt out conditions. As mentioned previously, this certainly could impact classification results, where we essentially exclude many teachers with high average prior achievement, which changes the relative rank of other teachers across the range.

CHAPTER 5

DISCUSSION

The purpose of this study was to investigate the extent to which student opt out of state assessments used for accountability impacts value-added measures of teacher effectiveness. As mentioned in Chapter 1, there has been a substantial increase in the number and proportion of students choosing to forego assessments administered in some states, the reasons for which appear to vary across locales.

No studies to date have asked students directly why they choose to opt out of assessments, but Pizmony-Levy and Green Saraisky (2016) surveyed opt out activists, most of whom were parents or teachers. Their study found that most activists advocated for students with which they had a personal relationship (i.e. their own children or children they taught) because they: a) disagreed with the implementation of Common Core State Standards, or b) disagreed with the use of assessments aligned to the Common Core for teacher or principal evaluations. A majority of the activists opted their children out of the assessments administered in their states, with 63.3 percent reporting they did and 93.1 reporting they were "very likely" or "likely" to do so in the future when they have a child in a tested grade. The authors also found that activists tended to be white women who were relatively wealthier than average, which may be representative of the students also choosing to opt out but actually represents a specific subset of the test-taking population. In New York, the State Education Department found that students wh chose to opt out were from wealthier districts, and were slightly more likely to be lower achieving than students who chose to take the assessment. Rice, Marland, and Meyer (2016) found corroborating results in New York, and added that these opt out students in

districts they studied were also slightly more likely to require special educational services. In Oregon, opt out students were reported to be wealthier, higher-achieving students (Hammond, 2009).

Given the demographic trends of this phenomenon, it is fair to say that opting out is potentially nonrandom, and that students who are no longer included in the test-taking population are systematically different. Accepting these facts, one could hypothesize that excluding these students from accountability measures (both achievement and growth) could potentially affect calculations and the resulting inferences about schools and educators. This study specifically focuses on the extent to which growth measures, as implemented in a value-added model and used for educator accountability, are impacted by nonrandom opt out trends in various magnitudes. In particular, the questions I set out to answer were:

1. What is the impact of opt out on value-added measures?
    a. How does opt out in different magnitudes within a teacher's classroom impact value-added measures?
    b. How do varying degrees of relationship between opt out patterns and prior achievement impact value-added measures?
2. What is the impact of opt out on classification of teachers value-added estimates using realistic classification systems?

This chapter discusses the results from the previous chapter in greater detail. Information is presented for each of the four simulated opt out conditions separately. First, results for the random condition are presented, including checks on the data generation process. Results for the highest probability condition are discussed next, then

for the highest achieving condition, and finally, for the lowest achieving condition. Within each of the opt out conditions, the impact of varying magnitudes of opt out is considered as well. Then, a summary of each of the three nonrandom conditions compared with the random condition will be discussed. Finally, implications of the results for evaluating teachers, as well as the limitations of this particular study design, is discussed.

5.1 Random Opt Out

The random opt out condition was created primarily as a sensitivity check, which allowed me to compare the extent to which the simulated nonrandom conditions present additional challenges not seen with random opt out. As mentioned, a large degree of random opt out could potentially impact the number of teachers receiving value-added estimates, as well as potentially impacting the estimates, standard errors, and the classification of teachers into effectiveness categories. While I consider random opt out in this section, a comparison to random opt out results will also be included in each of the following three conditions as well.

In general, results from the random opt out condition were expected, and were perhaps slightly better than expected. In terms of the data generation process, the percent of random opt out was generally in the magnitude category being simulated, and the average prior achievement was the same for students chosen for opt out and for those who were not. Correlations between average prior achievement and opting out were zero at both the student and teacher level.

The fixed effect distributions were normal, with a mean of 0, a standard deviation of .22, and kurtosis of approximately 3. Correlations between value-added estimates from

the complete condition and each of the random magnitude conditions (5, 10, and 20 percent) were at or above .983. The RMSD's of value-added estimates were lowest in this condition, ranging from .019 in the 5 percent magnitude condition to .04 in the 20 percent magnitude condition. RMSD of standard errors were also lowest, with a range of .003 in the 5 percent condition to .011 in the 20 percent condition. Classification agreement was fairly high for the random condition, at 91.1 percent in the 5 percent condition and 80.6 in the 20 percent. Finally, classification agreement was also relatively similar across each of the prior achievement quartiles.

We can interpret these results to mean that teachers might expect to see slight differences in their value-added estimates in states where opt out occurs totally at random. In the worst-case scenario (for instance) an average teacher might see his or her value-added estimate change by about .18 of a standard deviation if 20 percent of students opted out randomly. In addition, using the classification scheme implemented here, almost 20 percent of teachers would move up or down one quartile in the classification process. Finally, the magnitude of opt out appears to have a substantial impact on each of the stability statistics presented here, with higher RMSD's and lower classification agreement as the magnitude of opt out increases.

5.2 Highest Probability Opt Out

As mentioned, the relationship between student prior achievement and opting out was relatively weak in the year used to generate data for this study. The correlation at the student level between prior achievement and opting out in New York State was approximately -0.10, and -0.05 at the teacher level. However, the nature of the relationship appeared to be nonrandom because the nonzero correlations were negative at

both levels. These correlations were used to generate the opt out indicators for the highest probability condition – potentially the most realistic condition of the four simulated ones.

In the highest probability condition, the percent of opt out in teachers' classrooms was also similar to the magnitude being generated. For instance, in the 5 percent condition, 48,561 (or 48.6 percent) had between 0 and 10 percent opt out. Average prior achievement was about 6 to 7 points lower for opt out students than it was for students who did not opt out, which we expect due to the specification of negative correlations. Correlations between prior achievement and whether a student opted out were between -0.05 and -0.07 at the student level, and -0.04 and -0.05 at the teacher level. Correlations between value-added and the percent opt out were slightly negative at -0.01 for all magnitude conditions.

Distributional characteristics of value-added estimates were the same as for all the other conditions, with a mean of zero and a standard deviation of about .22. Correlations between the complete value-added and each of the three magnitudes ranged from .996 in the 5 percent condition down to .981 in the 20 percent condition. RMSD's of the value-added estimates and standard errors were similar to that of the random condition, though slightly higher by approximately .001. Finally, the classification agreement was just slightly lower than in the random condition, with 91 percent agreement in the 5 percent condition and 80 percent in the 20 percent condition.

In this study, the highest probability condition mirrored the results of the random condition, perhaps because of the relatively weak negative correlations used in the data generation process. However, these correlations existed in the empirical data, and so must be studied to determine the extent to which they pose a threat to the inferences made in

New York State about teachers. Similar to the random condition, the magnitude of opt out did have a substantial impact on the the stability statistics used in this study.

5.3 Highest Achieving Condition

As of this writing, no other states had provided the same level of information about opt out as was provided by New York State. Because of this lack of information, two additional conditions, perhaps more extreme than actual situations, were created to simulate other possible scenarios. The first scenario is the highest achieving condition, where 50 percent of students selected for opt out were from classrooms in the top quartile of prior achievement. The remaining 50 percent were selected from the other three quartiles of prior achievement at random. This condition led to some interesting distributional characteristics in the data generation process, as well as additional challenges to the use of value-added estimates.

In the highest achieving condition, there were substantially more teachers with higher levels of opt out for each of the magnitudes when compared with the random condition. For instance, more than 12,000 teachers had between 25 and 50 percent opt out, and more than 13,000 teachers had more than 50 percent. This phenomenon is likely due to the fact that I performed a very targeted selection of opt out students in the 250 classrooms with the highest prior achievement in each simulation. This also led to the highest number of teachers being completely excluded from analysis, with more than 10,000 dropping out because they had 10 or fewer students included in their value-added estimate. Average prior achievement was substantially higher for opt out students, when compared with students who did not opt out. The correlation at the student level between

prior achievement and opting out ranged from .31 to .39, and between .64 and .74 at the teacher level.

There was slightly more variability in the fixed effects for the highest achieving condition, with standard deviations ranging from .23 to .25 (compared with .22 in other conditions). Correlations between the complete and incomplete value-added at each of the three magnitudes ranged from .995 in the 5 percent condition to .982 in the 20 percent condition. RMSD's of value-added estimates were highest for this condition, ranging from .025 to .05. RMSD's of the standard errors were also double that of the random condition, and ranged from .008 to .022. Classification agreement was approximately similar to the random condition for the 5 and 10 percent magnitude, but was about 3 percentage points lower in the 20 percent condition.

Upon further analyses, we see that teachers in the top quartile of prior achievement were the most likely to change effectiveness quartiles. The average percent of top quartile teachers changing one quartile was 12 percent in the 5 percent condition, and as high as 26.5 percent in the 20 percent condition. This is likely due to the fact that most teachers who ultimately get excluded from analysis are those in the top quartile of prior achievement, which changes the relative ranking of other teachers.

The highest achieving condition had what one might consider the most extreme results, with the highest RMSD's of standard errors across replications, as well as the lowest classification agreement. As mentioned, this was meant to simulate a situation in which the highest achieving students from high achieving classrooms choose to opt out of the assessment. The actual implementation of this may not be totally realistic, but it does

provide some information about how value-added estimates and classification could change as a result of nonrandom opt out of the highest achieving students.

5.4 Lowest Achieving Condition

The final simulation was one where the lowest achieving students in the highest achieving classrooms choose to opt out of the assessment. Similar to the highest achieving condition, 50 percent of opt out students are the lowest achieving in classrooms in the top quartile of prior achievement, while the other 50 percent were chosen at random across the other three quartiles.

In terms of the percent of opt out, we saw very similar results to the highest achieving condition, with substantially more classrooms with more than 50 percent opt out in the 20 percent condition. More than 9,000 teachers were excluded from stability analyses because they had 10 or fewer students included in their value-added estimate. Average prior achievement was lower for opt out students than for non-opt out students, which we would expect. However, the difference between the opt out and non-opt out students was not as large in the lowest achieving condition as it was in the highest achieving condition. This may be due to the fact that I am selecting the lowest achieving students in the top quartile of classrooms, which are not necessarily the lowest achieving students in the entire distribution.

Correlations between prior achievement and opting out ranged from -0.14 to -0.06 at the student level, and from .45 to .6 at the teacher level. The change in the direction of the correlation was meant to mirror what Rice, Marland and Meyer (2016) found in their analyses of New York State where the direction changed from student to school or district level. The distributional characteristics were similar to those of the other three opt

out conditions, with a mean of zero and standard deviation of .22, and correlations between complete and incomplete value-added estimates at each magnitude were also similar to the other opt out conditions. RMSD of value-added estimates and standard errors were also similar to the random and highest probability condition, which were lower than what was found in the highest achieving condition.

Classification agreement in the lowest achieving condition was similar to what was found in the highest achieving condition, ranging from 78.6 in the 20 percent condition to 91.3 in the 5 percent condition. Also similar to the highest achieving condition, agreement was lowest for the classrooms in the top quartile of prior achievement. On average, 11.7 of teachers in the top quartile could be expected to change one quartile in the 5 percent condition, while 26.3 percent of teachers in the top quartile could be expected to change in the 20 percent condition.

5.5 Implications of Findings

There are two prominent findings that can help to explain the changes in classifications for teachers. The first is that results across opt out conditions were relatively similar, with the exception of the 20 percent condition where the standard errors were larger and classification agreement lower than for the other three conditions. Beyond that, distributional characteristics, correlations between complete and incomplete value-added estimates, RMSD's of the value-added estimates, and even classification agreement were, for the most part, fairly similar.

However, the magnitude of opt out did appear to have a large impact on stability statistics, where we saw that RMSD's of value-added estimates doubled when opt out increases from 5 to 20 percent of students choosing to opt out, and nearly tripled for

standard errors. Classification agreement dropped 11 - 14 percent from approximately 91 percent to as low as 77 percent across opt out conditions as well.

Comparing the results of the lowest achieving condition to each of the other three conditions helps to better understand what actually causes the change to classification for teachers. The RMSD of the value-added estimates and standard errors for the lowest achieving condition were similar to the highest probability and random conditions, while the percent opt out and classification agreement are more similar to the highest achieving condition.

In addition, Figures 32 to 39 make two additional points that help to complete the picture. First, Figures 32 to 35 shows that there is a strong relationship between value-added estimates, and that those who change quartiles are closer to zero. Second, Figures 36 to 39 shows that changes in quartiles were relatively uniform across average prior achievement up until the 325 score on the achievement scale. At this point, we see a substantial increase in the number of teachers who were excluded from analysis completely.

All of this evidence taken together can be interpreted to mean that the magnitude of opt out, in the ways that are simulated, is what causes the change in teacher classification. A far more substantial portion of teachers was completely excluded from classification in the highest and lowest achieving conditions, which likely caused the remaining teachers to change their relative ranking in the distribution of fixed effects.

5.6 Limitations

This is a simulation study, which carries with it some limitations regarding generalization to realistic settings. I used empirical data from one state to calculate

parameter estimates for the data generation process, but only those parameters were controlled in the process. I used the mean scale scores from empirical data, as well across-year correlations of scale scores at the student and teacher level, student and teacher-level scale score standard deviations, the intraclass correlation (ICC) of opt out and scale scores, and the correlation between prior achievement and opting out at the teacher and student levels. While this certainly represents many important aspects of generating nested classroom scale score data, there are some factors that were not controlled. For instance, students in realistic settings are affected by grade-, school- and district-level influences as well, which were not included in this study. The ICCs of scale scores and percent of opt out were used to represent between classroom differences that suggest non-random assignment of students to teachers, but the teacher-level ICC neglects school and district-level differences in achievement. Including school and district effects might also create more variation in teacher fixed effects generated as part of this study, and should be considered in the next round of analysis.

In addition, the empirical data used to generate the parameter estimates were a sub-sample of the state, and did not fully represent the state as well. Only 37.6 percent of students were considered as living in poverty in my sample, where 51.9 percent were statewide. This fact may affect the generated parameter estimates somewhat, if students living poverty tend to have different growth trajectories than those who do not (which is the case in other locales).

Relatedly, as of this writing, no other states have published parameter estimates from which I could conduct a similar study to what was done for New York. Because of the lack of information, the highest and lowest achieving conditions are somewhat

manufactured. These two conditions were meant to mirror situations that were conceptually similar, but it is impossible to know whether the conditions I created are close to, or far from, the realities in those locales.

In addition, the standard deviation of value-added estimates were approximately .22 - .24, where in empirical data they tend to be between .25 - .35. As mentioned, this may be due, in part, to not including between-school and district differences in achievement in the data generation process.

5.7 Conclusion

This study has several implications for states and districts where opt out is non-negligible for several reasons. The first is that states put these growth measures in place to hold teachers, schools, and districts accountable for improving student learning, and ultimately, many are not accountable because of a reduced number of students eligible for inclusion in value-added estimates – while at the same time, they may have encouraged students to opt out from the assessment. This brings forward the second implication.

The choice by these educators impacts their colleagues as well if states employ a normative classification scheme, where the place in the distribution of fixed effects ultimately determines a teacher's classification. If opt out trends continue, states and districts may find a criterion-referenced classification preferable, where value-added thresholds are used to classify teachers into effectiveness categories. As seen in the literature, this requires experts to determine what qualifies as low, average, and high growth. However, a criterion-reference system may be preferable to a system where teachers are more likely to change classifications because of another teacher's behavior.

States and districts should consider standard errors when classifying teachers into effectiveness categories. As seen in the analyses, standard errors increase substantially when there are large degrees of opt out in a teacher's classroom, which diminishes one's ability to interpret the estimate with any degree of confidence. Some states, including New York, already use a confidence interval when classifying teachers. However, they also use a normative approach to classification.

Opt out in large magnitudes, like in New York and other states, causes downstream effects related to data availability as well. With respect to achievement measures, students who are excluded from analysis this year cannot be included in calculating the percent of students meeting the grade-level proficiency standards. While the percent meeting grade-level standards may be an accurate measure of those included in the analysis, it may not be accurate for all students in a school or district if opt out students represent a systematically different sample. That is to say that the percent may be higher or lower if opt out students participated in the assessment. In addition, in subsequent years, calculating adequate yearly progress could be biased if some previous opt out students choose to participate, or if new students choose to do so in a nonrandom manner. For instance, if opt out becomes more nonrandom, the calculation of the change in the percent of students meeting standards may be incomparable because the two groups are no longer similar to each other.

With respect to growth measures, in addition to the issues mentioned above, one additional downstream effect is that opt out students this year likely cannot be included in growth measures next year. All growth models require a baseline score from which to calculate change, which becomes much more difficult for students who choose to opt out.

There is a future line of inquiry related to utilizing multiple imputation of missing prior year scores, or in using prior scores from two years prior as a baseline; however, both of those introduce a new set of issues. As mentioned, multiple imputation requires acknowledging variability due to sampling error, which may be problematic to those being evaluated using these scores. The two-year lag correlation is also slightly lower than a one-year lag, which introduces additional uncertainty into value-added estimation. Both methods assume that students who opt out this year eventually participate in assessments in future years, which also needs to be investigated. It is possible that students who choose to opt out this year will continue to do so in future years, or that they may return to the data. This question is easily answered, now that states have several years of data where opt out has been seen in non-negligible magnitudes.

States with large proportions of opt out, regardless of the type, should consider ways to adjust for the resulting differences in value-added estimates outlined here. The best case scenario for a state in this study was that 20 percent of teachers move one quartile when 20 percent of students opt out. Teachers who moved were primarily in the middle two quartiles, but it is possible that some could end up in a quartile of consequence. Given that, states might consider the amount of opt out in a teacher's classroom when creating a rating for them. For instance, a teacher with no opt out who moves into a quartile of consequence might be "held to no harm" as a result of the large degree of opt out elsewhere.

Table 1: Percent of Opt Out by State from Bennett (2016)

State	**ELA**	**Math**	**Overall**
California			3%
Idaho			2%
Connecticut			4%
Washington	2%	3%	
Maine	5%	6%	
Colorado	11%	10%	
Rhode Island	12%	10%	
New York			20%

Table 2: Number and Percent of Teachers with Varying Levels of Non-Participation on the State Assessment (NYSED, 2015)

	Number of Teachers	Percent of Teachers
Teachers with exactly 0% non-participation	7,005	19.59
Teachers with more than 0% and less than 10% non-participation	10,688	29.89
Teachers with 10% to 25% non-participation	6,551	18.32
Teachers with 25% to 50% non-participation	7,831	21.9
Teachers with 50% or more non-participation	3,680	10.29

Table 3:Descriptive Statistics for 6th Grade Math in Sample and Statewide

	Sample	Statewide
Number of Opt Out Students	1,563	47,177
Number of Participating Students	6,460	141,167
Total Enrolled	8,023	188,344
% of Opt Out	0.19	0.25
S.D. of Opting Out	0.4	n/a

Table 4: Number and Percent of Teachers in Each Opt Out Category by Magnitude: Random Condition

	Random					
	5%		10%		20%	
	N	%	N	%	N	%
0 percent	24,683	24.7	7,360	7.4	1,084	1.1
> 0 percent & <= 10 percent	64,170	64.2	47,539	47.5	8,809	8.8
> 10 percent & <= 25 percent	11,037	11.0	43,833	43.8	67,494	67.5
> 25 percent & <= 50 percent	103	0.1	1,235	1.2	22,524	22.5
> 50 percent	7	0.0	33	0.0	89	0.1

Table 5: Number and Percent of Teachers in Each Opt Out Category by Magnitude: Highest Probability Condition

	Highest Probability					
	5%		10%		20%	
	N	%	N	%	N	%
0 percent	36,014	36.0	16,203	16.2	4,151	4.2
> 0 percent & <= 10 percent	48,561	48.6	43,713	43.7	20,279	20.3
> 10 percent & <= 25 percent	14,436	14.4	33,972	34.0	46,055	46.1
> 25 percent & <= 50 percent	975	1.0	5,962	6.0	27,471	27.5
> 50 percent	14	0.0	150	0.2	2,044	2.0

Table 6: Number and Percent of Teachers in Each Opt Out Category by Magnitude: Highest Achieving Condition

	Highest Achieving					
	5%		10%		20%	
	N	%	N	%	N	%
0 percent	37,423	37.4	18,293	18.3	5,507	5.5
> 0 percent & <= 10 percent	44,253	44.3	49,523	49.5	35,495	35.5
> 10 percent & <= 25 percent	15,267	15.3	21,888	21.9	32,975	33.0
> 25 percent & <= 50 percent	2,943	2.9	9,504	9.5	12,248	12.2
> 50 percent	114	0.1	792	0.8	13,775	13.8

Table 7: Number and Percent of Teachers in Each Opt Out Category by Magnitude: Lowest Achieving Condition

	Lowest Achieving					
	5%		10%		20%	
	N	%	N	%	N	%
0 percent	37,729	37.7	18,428	18.4	5,526	5.5
> 0 percent & <= 10 percent	42,656	42.7	50,129	50.1	35,522	35.5
> 10 percent & <= 25 percent	17,566	17.6	18,645	18.6	33,822	33.8
> 25 percent & <= 50 percent	2,044	2.0	12,765	12.8	10,860	10.9
> 50 percent	5	0.0	33	0.0	14,270	14.3

Table 8: Number and Percent of Teachers with Exactly Zero Students in Each Condition and Magnitude

	Random		Highest Probability		Highest Achieving		Lowest Achieving	
Magnitude	N	%	N	%	N	%	N	%
5%	5	0.005	5	0.005	11	0.011	4	0.004
10%	25	0.025	29	0.030	43	0.040	9	0.009
20%	5	0.005	5	0.005	37	0.037	6	0.006

Table 9: Number and Percent of Teachers with 10 or Fewer Students Included in Value-Added Estimates in Each Condition and Magnitude

	Complete VA		Random		Highest Probability		Highest Achieving		Lowest Achieving	
Magnitude	N	%	N	%	N	%	N	%	N	%
5%	1,900	1.9	2,687	2.7	2,734	2.7	2,941	2.9	2,822	2.8
10%	2,300	2.3	3,144	3.1	3,318	3.3	3,631	3.6	3,461	3.5
20%	1,900	1.9	5,229	5.2	5,980	6.0	10,551	10.6	9,349	9.3

Table 10: Average Student-Level Prior Achievement by Opt Out Status, Opt Out Condition, and Magnitude

		Average Prior Achievement	
Condition	Magnitude	Opt Out	Not Opt Out
Random	5%	310	310
	10%	310	310
	20%	310	310
Highest Probability	5%	303	310
	10%	304	311
	20%	305	311
Highest Achieving	5%	355	307
	10%	346	306
	20%	336	303
Lowest Achieving	5%	290	311
	10%	296	311
	20%	306	311

Table 11: Student- and Teacher-Level Correlations Between Prior Achievement/Value-Added and Opt Out

		Random Opt Out		Highest Probability		Highest Achieving		Lowest Achieving	
		Student	Teacher	Student	Teacher	Student	Teacher	Student	Teacher
Prior Achievement	5%	0.00	0.00	-0.05	-0.04	0.31	0.64	-0.14	0.45
	10%	0.00	0.00	-0.06	-0.04	0.34	0.68	-0.13	0.50
	20%	0.00	0.00	-0.07	-0.05	0.39	0.74	-0.06	0.60
Value-Added	5%		0.00		-0.01		-0.01		-0.01
	10%		0.00		-0.01		-0.02		-0.02
	20%		0.00		-0.01		-0.02		-0.02

Table 12: Distributional Descriptive Statistics for Value-Added Estimates by Condition and Magnitude

		Complete	Random Opt Out	Highest Probability	Highest Achieving	Lowest Achieving
Mean	5%	0.00	0.00	0.00	0.00	0.00
	10%	0.00	0.00	0.00	0.00	0.00
	20%	0.00	0.00	0.00	0.00	0.00
Standard Deviation	5%	0.22	0.22	0.22	0.23	0.22
	10%	0.22	0.22	0.22	0.24	0.22
	20%	0.22	0.23	0.23	0.25	0.22
Kurtosis	5%	3.02	3.02	3.02	3.03	3.02
	10%	2.99	2.99	2.99	2.99	2.99
	20%	3.01	3.00	3.01	3.00	3.01

Table 13: Correlations Between Complete and Incomplete Value-Added Estimates from Each Condition and Magnitude

	5%	10%	20%
Random	0.996	0.992	0.983
Highest Probability	0.996	0.991	0.981
Highest Achieving	0.995	0.991	0.982
Lowest Achieving	0.996	0.991	0.981

Table 14: Stability Statistics for Value-Added Estimates for Each Condition and Magnitude

		Random	Highest Probability	Highest Achieving	Lowest Achieving
RMSD - VA	5%	0.019	0.020	0.025	0.021
	10%	0.028	0.029	0.035	0.029
	20%	0.042	0.043	0.050	0.043
RMSD - SE	5%	0.003	0.004	0.008	0.004
	10%	0.005	0.006	0.013	0.007
	20%	0.011	0.012	0.022	0.014
Classification Agreement	5%	91.1	91.0	91.1	91.3
	10%	87.4	87.3	87.8	88.0
	20%	80.6	80.0	77.7	78.6

Table 15:Average Number and Percent of Teachers Who Change Rating Categories by Prior Achievement: Random Condition

		Random							
	Prior Achievement	No Change		+/-1 Category		+/-2 Categories		+/-3 Categories	
Magnitude	Quartile	N	%	N	%	N	%	N	%
5%	1 - Bottom	226.91	93.7	15.25	6.3	0.00	0.0	0.00	0.0
	2	228.20	93.5	15.85	6.5	0.00	0.0	0.00	0.0
	3	228.56	93.7	15.34	6.3	0.00	0.0	0.00	0.0
	4 - Top	227.62	93.7	15.40	6.3	0.00	0.0	0.00	0.0
10%	1 - Bottom	217.66	90.2	23.64	9.8	0.00	0.0	0.00	0.0
	2	220.09	90.5	23.02	9.5	0.00	0.0	0.00	0.0
	3	218.71	90.0	24.24	10.0	0.02	0.0	0.00	0.0
	4 - Top	217.69	90.3	23.49	9.7	0.00	0.0	0.00	0.0
20%	1 - Bottom	200.03	84.8	35.85	15.2	0.03	0.0	0.00	0.0
	2	201.92	84.9	35.97	15.1	0.03	0.0	0.00	0.0
	3	202.70	85.2	35.18	14.8	0.02	0.0	0.00	0.0
	4 - Top	201.69	85.5	34.27	14.5	0.02	0.0	0.00	0.0

Table 16: Average Number and Percent of Teachers Who Change Rating Categories by Prior Achievement: Highest Probability Condition

		Highest Probability							
		No Change		+/-1 Category		+/-2 Categories		+/-3 Categories	
Magnitude	Prior Achievement Quartile	N	%	N	%	N	%	N	%
5%	1 - Bottom	225.38	93.1	16.58	6.9	0.00	0.0	0.00	0.0
	2	227.43	93.3	16.21	6.7	0.01	0.0	0.00	0.0
	3	228.28	93.6	15.69	6.4	0.03	0.0	0.00	0.0
	4 - Top	228.67	94.1	14.38	5.9	0.00	0.0	0.00	0.0
10%	1 - Bottom	216.12	89.8	24.48	10.2	0.04	0.0	0.00	0.0
	2	219.29	90.3	23.42	9.6	0.06	0.0	0.00	0.0
	3	219.37	90.4	23.42	9.6	0.00	0.0	0.00	0.0
	4 - Top	218.03	90.6	22.54	9.4	0.05	0.0	0.00	0.0
20%	1 - Bottom	196.91	84.5	35.99	15.4	0.15	0.1	0.00	0.0
	2	200.97	85.1	34.97	14.8	0.15	0.1	0.00	0.0
	3	201.52	85.3	34.64	14.7	0.07	0.0	0.00	0.0
	4 - Top	200.60	85.4	34.14	14.5	0.09	0.0	0.00	0.0

Table 17: Average Number and Percent of Teachers Who Change Rating Categories by Prior Achievement: Highest Achieving Condition

		Highest Achieving							
		No Change		+/-1 Category		+/-2 Categories		+/-3 Categories	
Magnitude	Prior Achievement Quartile	N	%	N	%	N	%	N	%
5%	1 - Bottom	233.38	95.9	9.86	4.1	0.00	0.0	0.00	0.0
	2	234.81	95.9	10.12	4.1	0.00	0.0	0.00	0.0
	3	234.68	95.8	10.18	4.2	0.00	0.0	0.00	0.0
	4 - Top	208.56	87.8	28.94	12.2	0.06	0.0	0.00	0.0
10%	1 - Bottom	226.85	93.6	15.56	6.4	0.00	0.0	0.00	0.0
	2	229.24	93.7	15.29	6.3	0.00	0.0	0.00	0.0
	3	228.32	93.4	16.02	6.6	0.00	0.0	0.00	0.0
	4 - Top	193.50	83.3	38.82	16.7	0.09	0.0	0.00	0.0
20%	1 - Bottom	217.60	90.2	23.76	9.8	0.00	0.0	0.00	0.0
	2	218.39	89.8	24.65	10.1	0.02	0.0	0.00	0.0
	3	218.51	90.1	23.91	9.9	0.01	0.0	0.00	0.0
	4 - Top	122.07	72.8	44.42	26.5	1.15	0.7	0.00	0.0

Table 18: Average Number and Percent of Teachers Who Change Rating Categories by Prior Achievement: Lowest Achieving Condition

		Lowest Achieving							
		No Change		+/-1 Category		+/-2 Categories		+/-3 Categories	
Magnitude	Prior Achievement Quartile	N	%	N	%	N	%	N	%
5%	1 - Bottom	233.22	95.9	10.05	4.1	0.00	0.0	0.00	0.0
	2	234.93	95.9	9.95	4.1	0.00	0.0	0.00	0.0
	3	234.51	95.8	10.36	4.2	0.00	0.0	0.00	0.0
	4 - Top	210.82	88.3	27.93	11.7	0.01	0.0	0.00	0.0
10%	1 - Bottom	227.49	93.8	15.07	6.2	0.00	0.0	0.00	0.0
	2	229.05	93.8	15.23	6.2	0.00	0.0	0.00	0.0
	3	228.17	93.4	16.01	6.6	0.00	0.0	0.00	0.0
	4 - Top	195.51	83.4	38.71	16.5	0.15	0.1	0.00	0.0
20%	1 - Bottom	217.27	90.1	23.85	9.9	0.02	0.0	0.00	0.0
	2	219.36	90.3	23.62	9.7	0.00	0.0	0.00	0.0
	3	217.55	90.1	23.96	9.9	0.02	0.0	0.00	0.0
	4 - Top	131.93	72.9	47.57	26.3	1.35	0.7	0.01	0.0

Figure 1: Missing Completely at Random

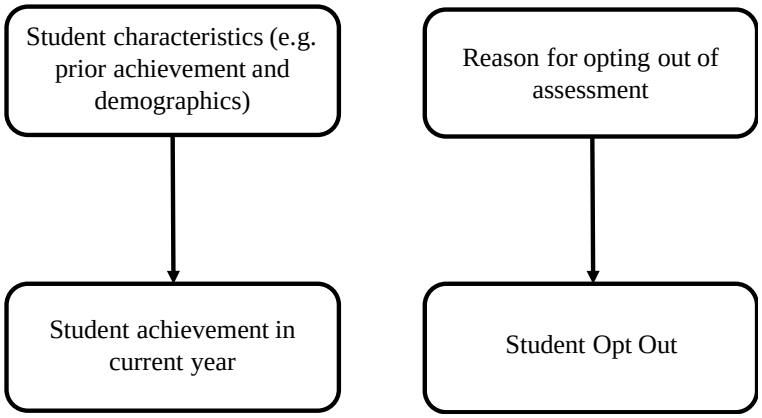


Figure 2: Missing at Random

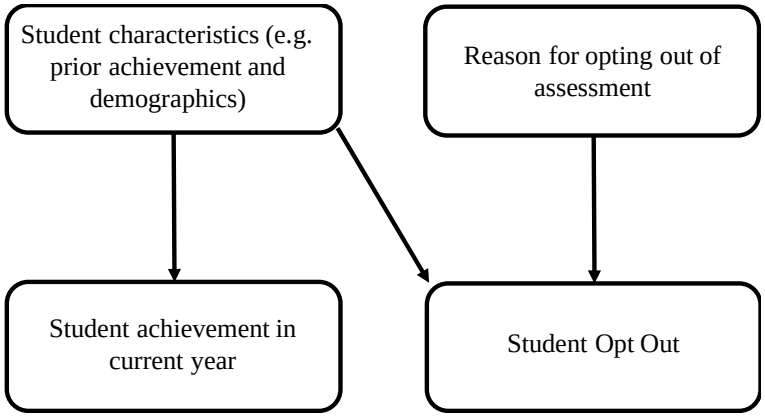


Figure 3: Missing Not at Random

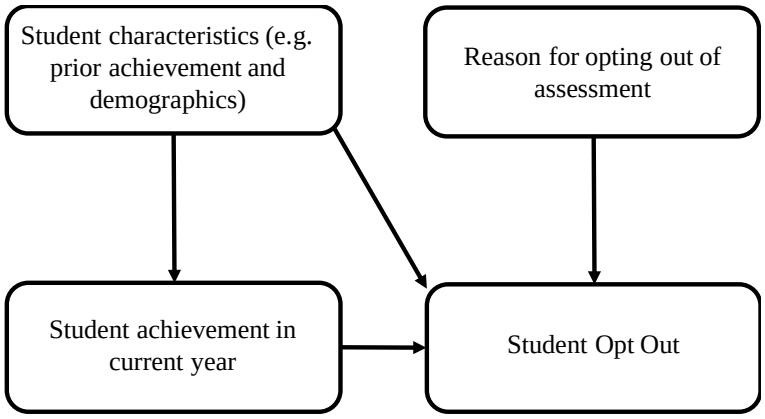


Figure 4:Distribution of Percent Opt Out in Each Classroom for Random Condition

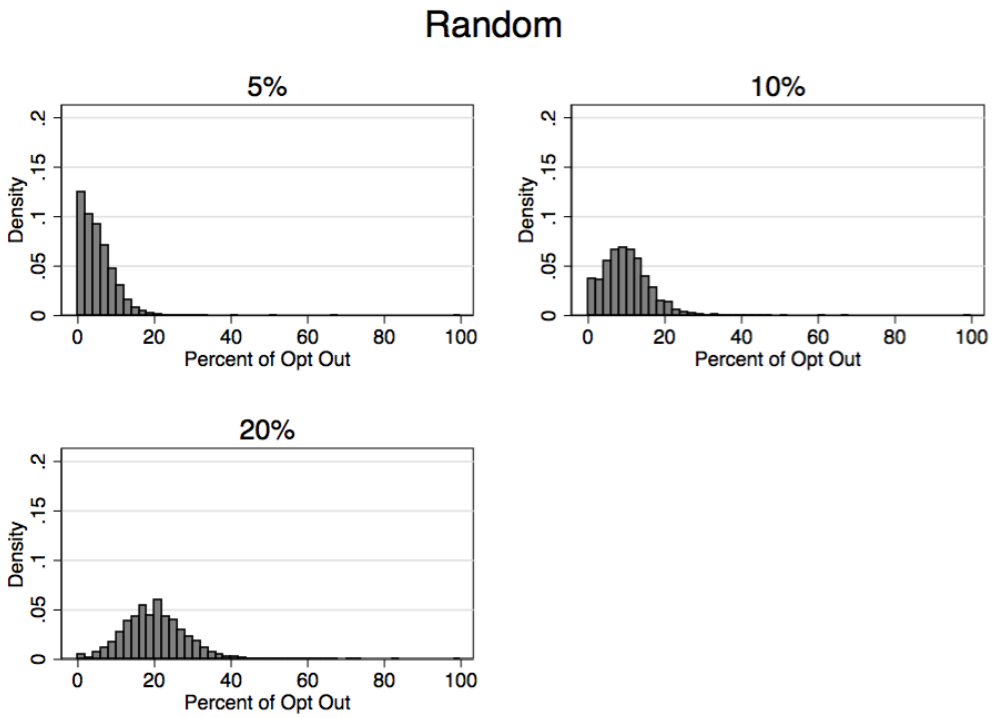


Figure 5: Distribution of Percent Opt Out in Each Classroom for Highest Probability Condition

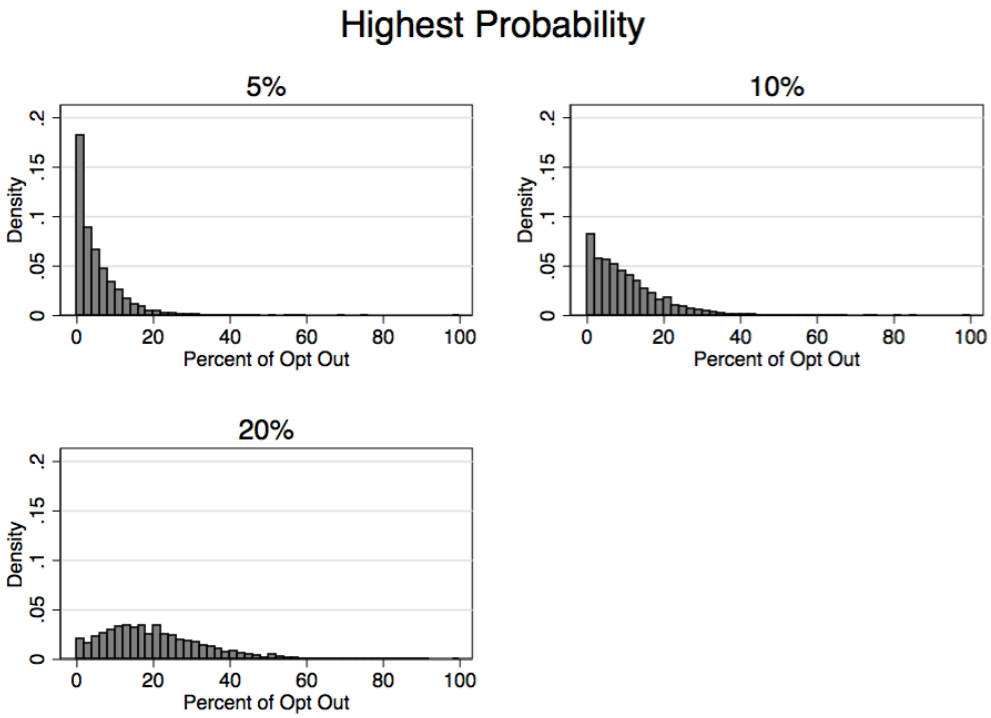


Figure 6: Distribution of Percent Opt Out in Each Classroom for Highest Achieving Condition

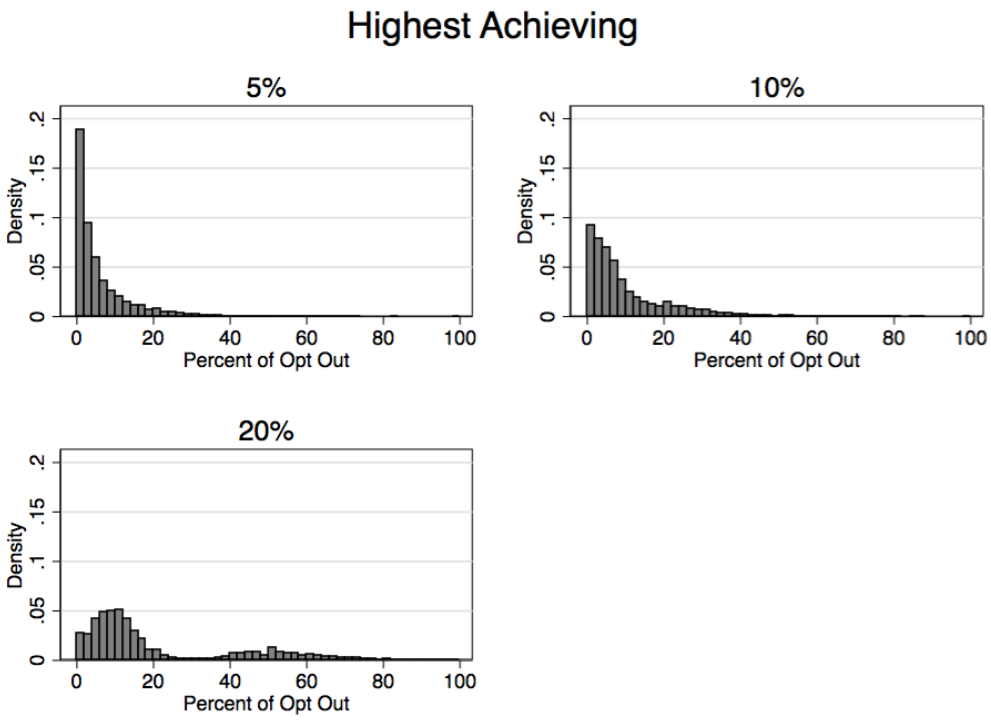


Figure 7: Distribution of Percent Opt Out in Each Classroom for Lowest Achieving Condition

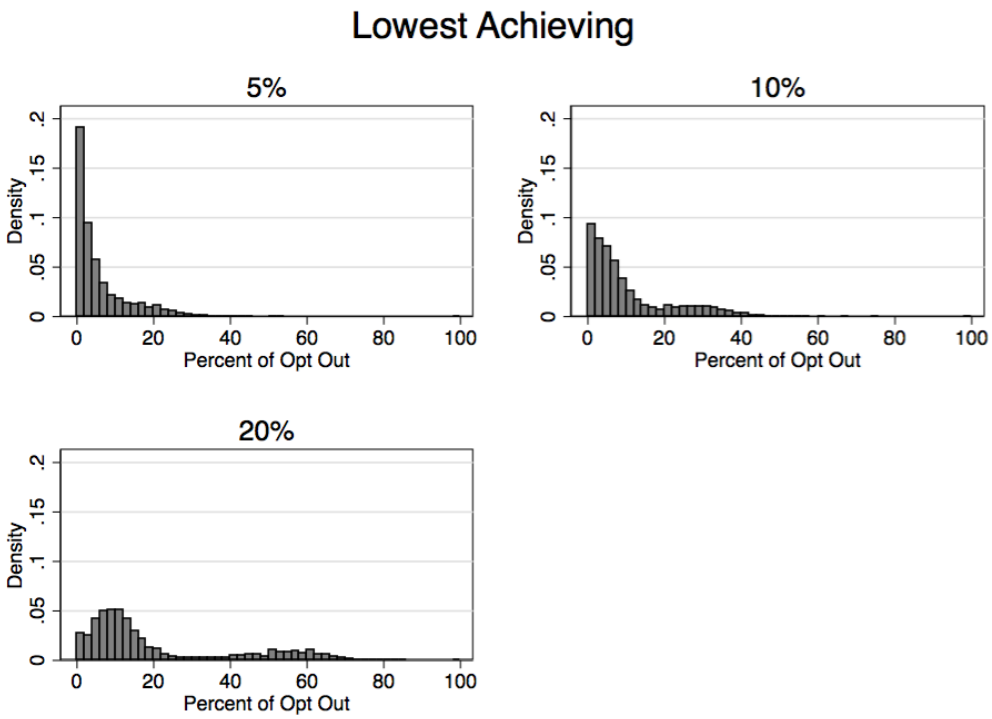


Figure 8: Distribution of Prior Achievement by Opt Out Status: Random Condition

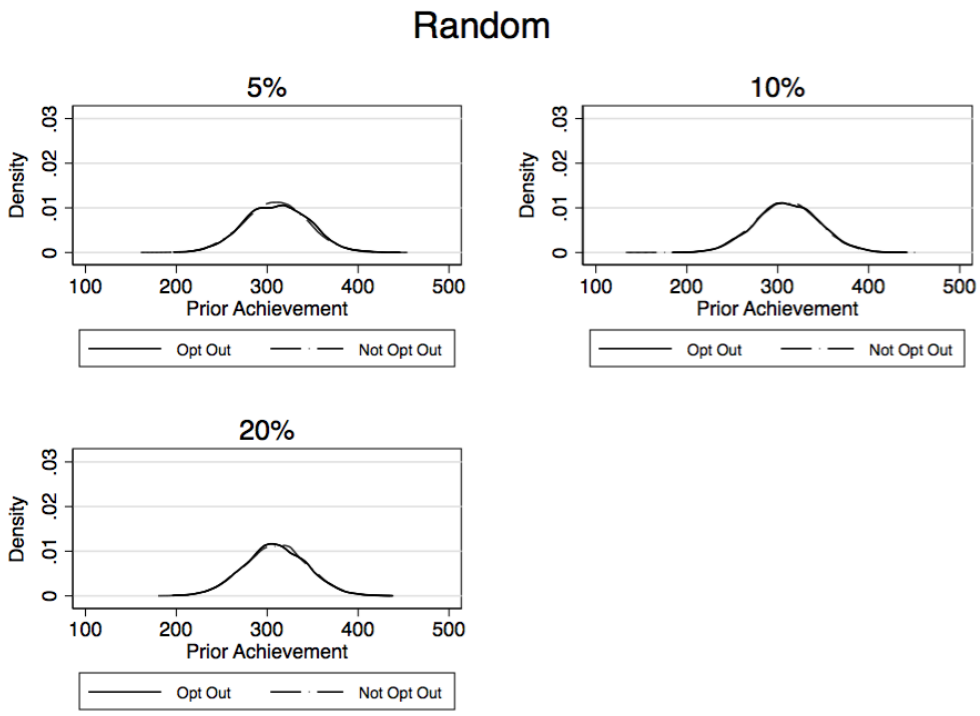


Figure 9: Distribution of Prior Achievement by Opt Out Status: Highest Probability Condition

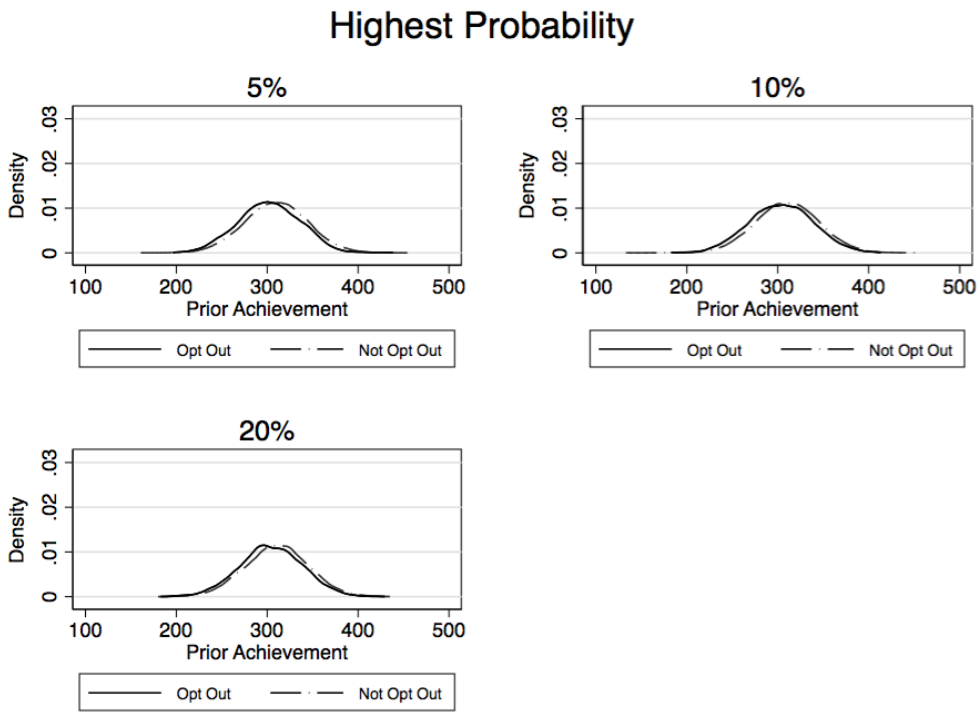


Figure 10: Distribution of Prior Achievement by Opt Out Status: Highest Achieving Condition

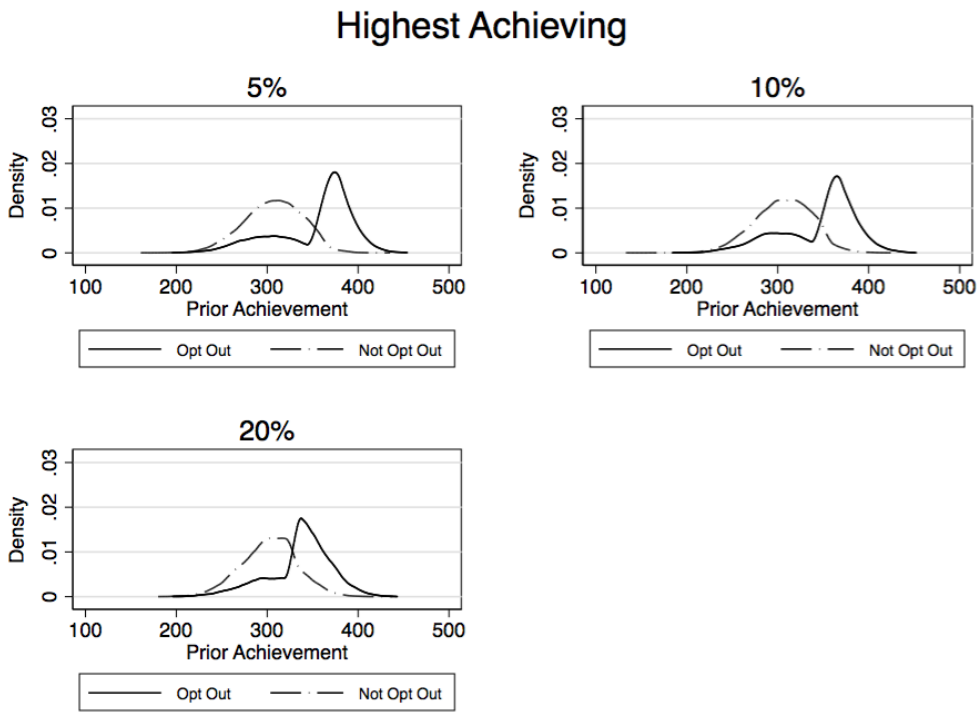


Figure 11: Distribution of Prior Achievement by Opt Out Status: Lowest Achieving Condition

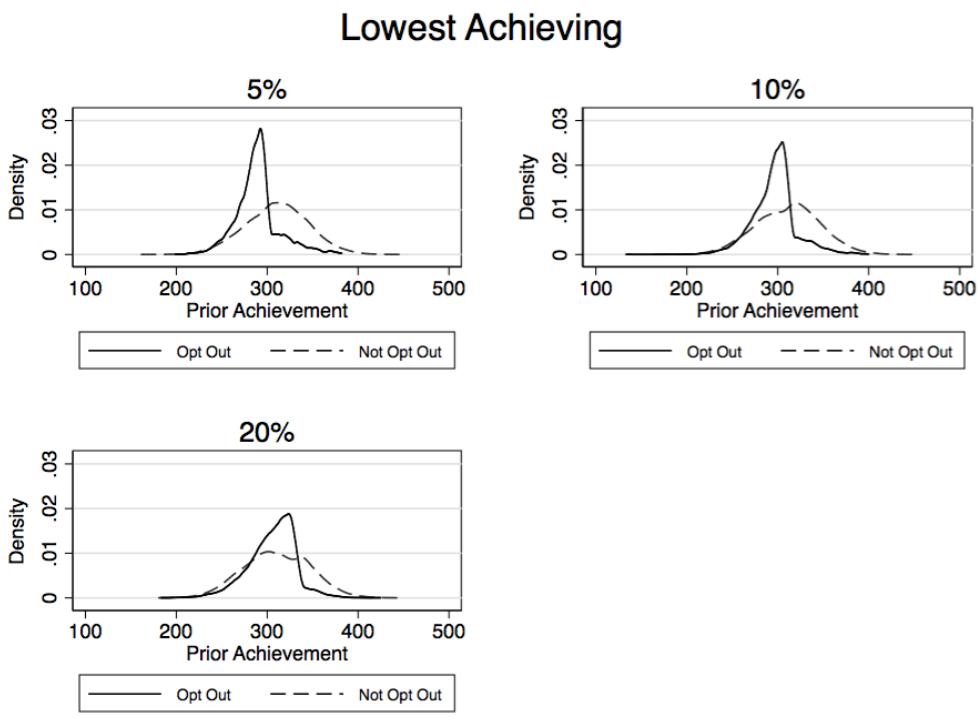


Figure 12: Percent of Opt Out by Average Prior Achievement: Random Condition (mspline smoothing, bands = 25)

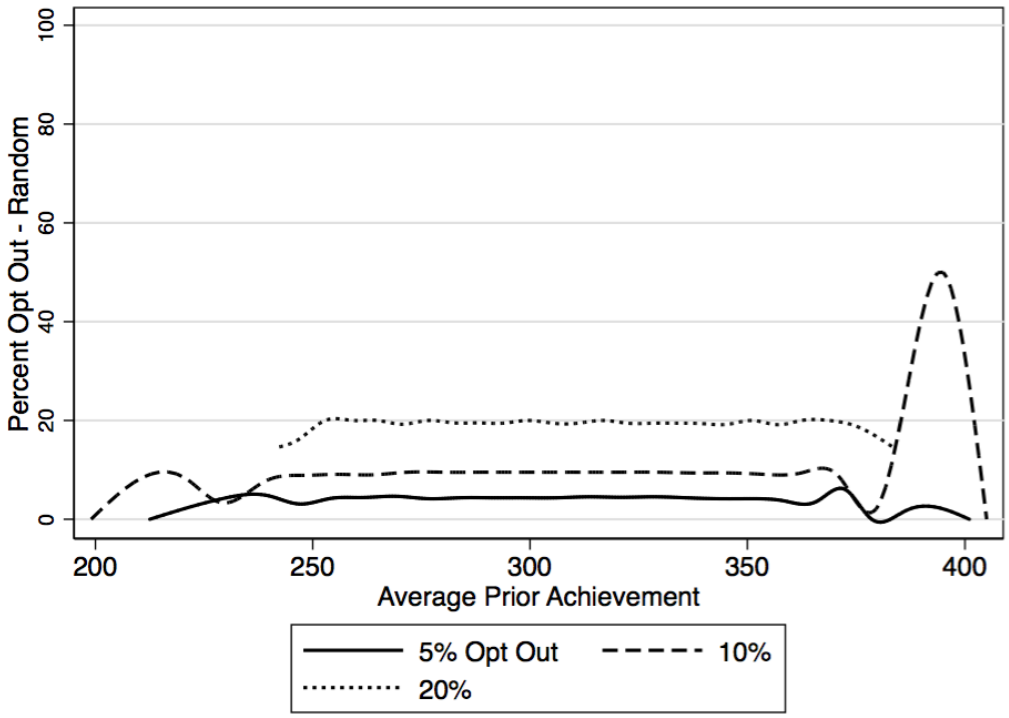


Figure 13: Percent of Opt Out by Average Prior Achievement: Highest Probability Condition (mspline smoothing, bands = 25)

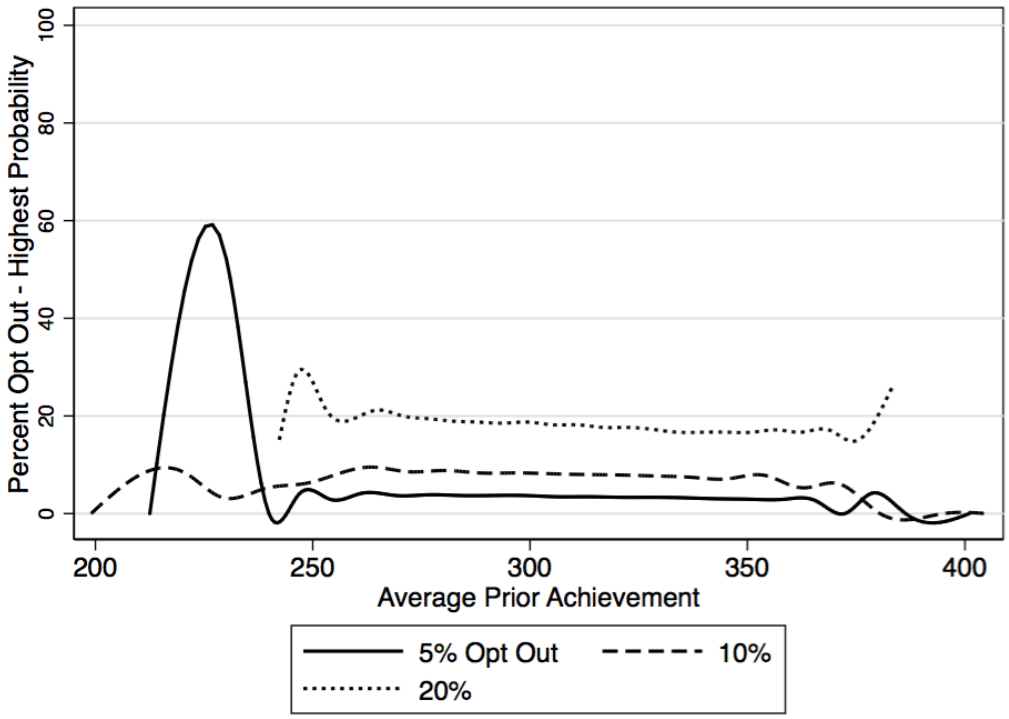


Figure 14: Percent of Opt Out by Average Prior Achievement: Highest Achieving Condition (mspline smoothing, bands = 25)

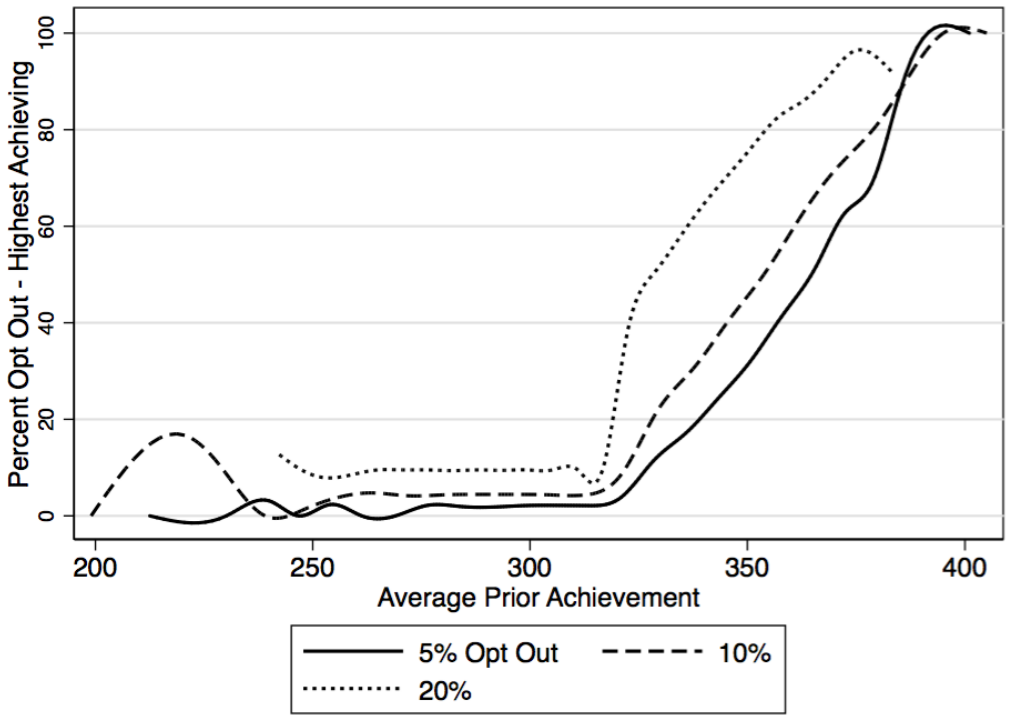


Figure 15: Percent of Opt Out by Average Prior Achievement: Lowest Achieving Condition (mspline smoothing, bands = 25)

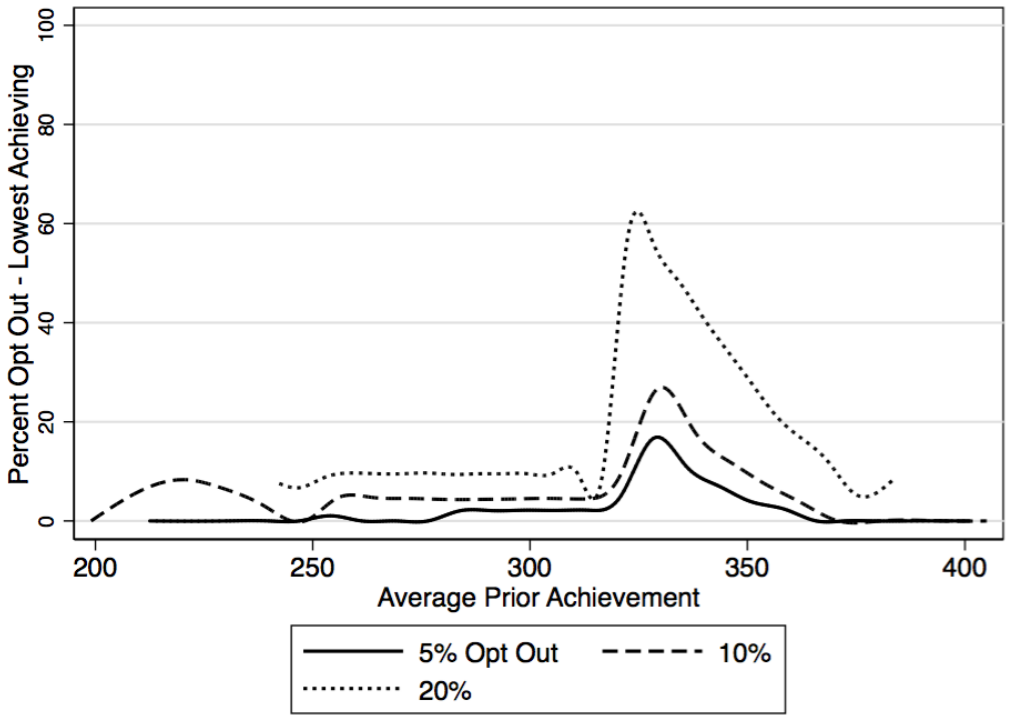


Figure 16: Difference in Complete and Incomplete VA Estimates by Prior Achievement: Random Condition (mspline smoothing, bands = 25)

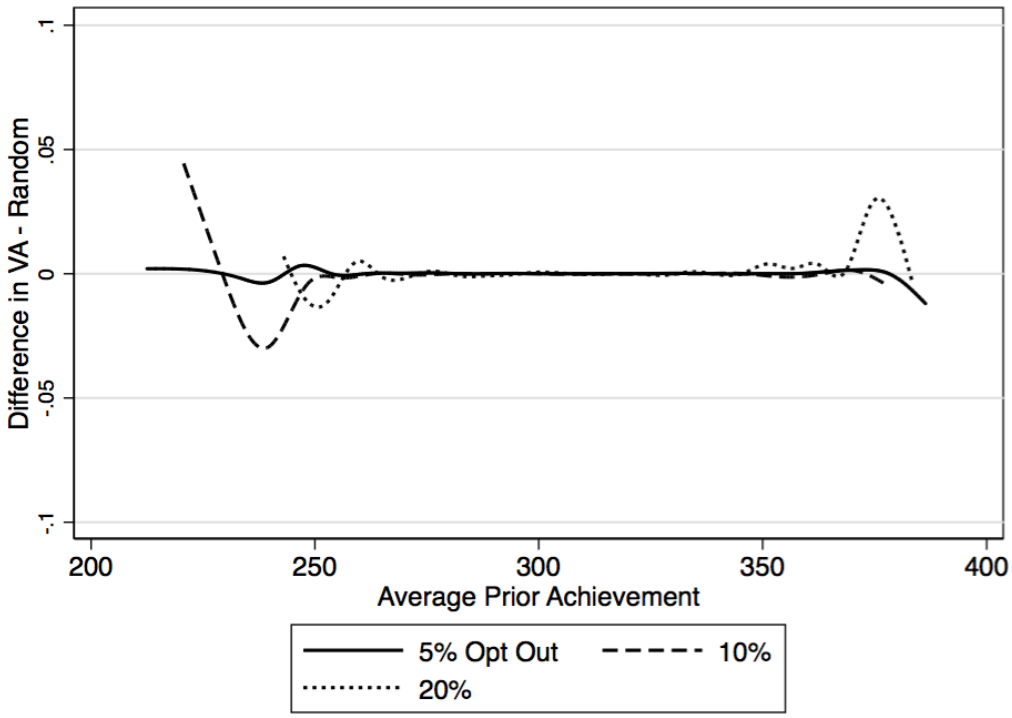


Figure 17: Difference in Complete and Incomplete VA Estimates by Prior Achievement: Highest Probability Condition (mspline smoothing, bands = 25)

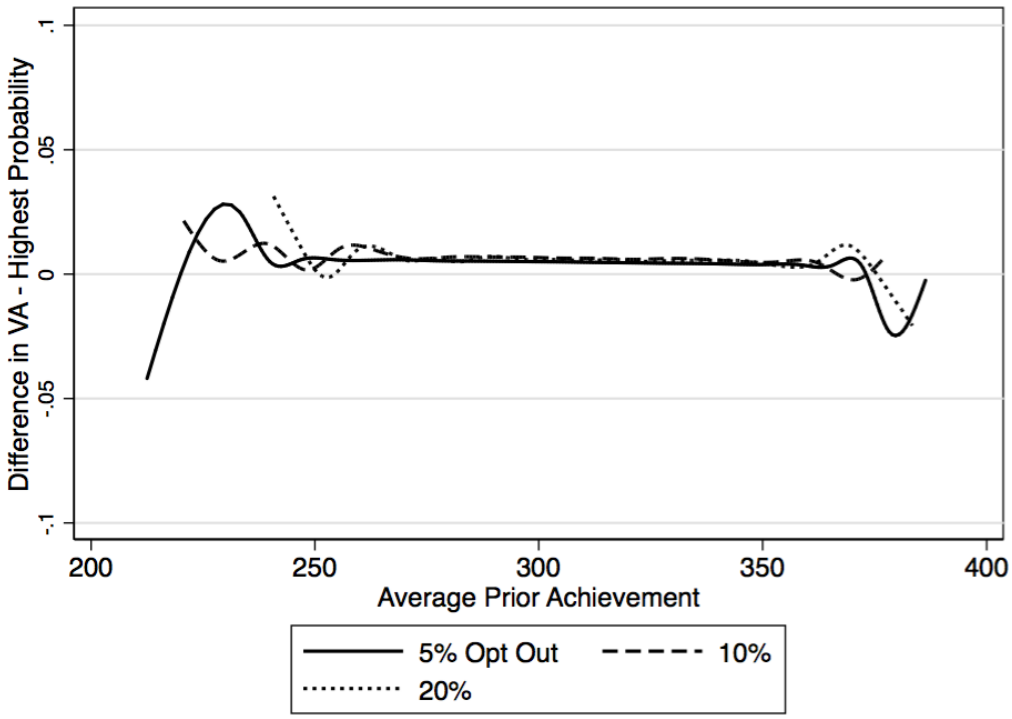


Figure 18: Difference in Complete and Incomplete VA Estimates by Prior Achievement: Highest Achieving Condition (mspline smoothing, bands = 25)

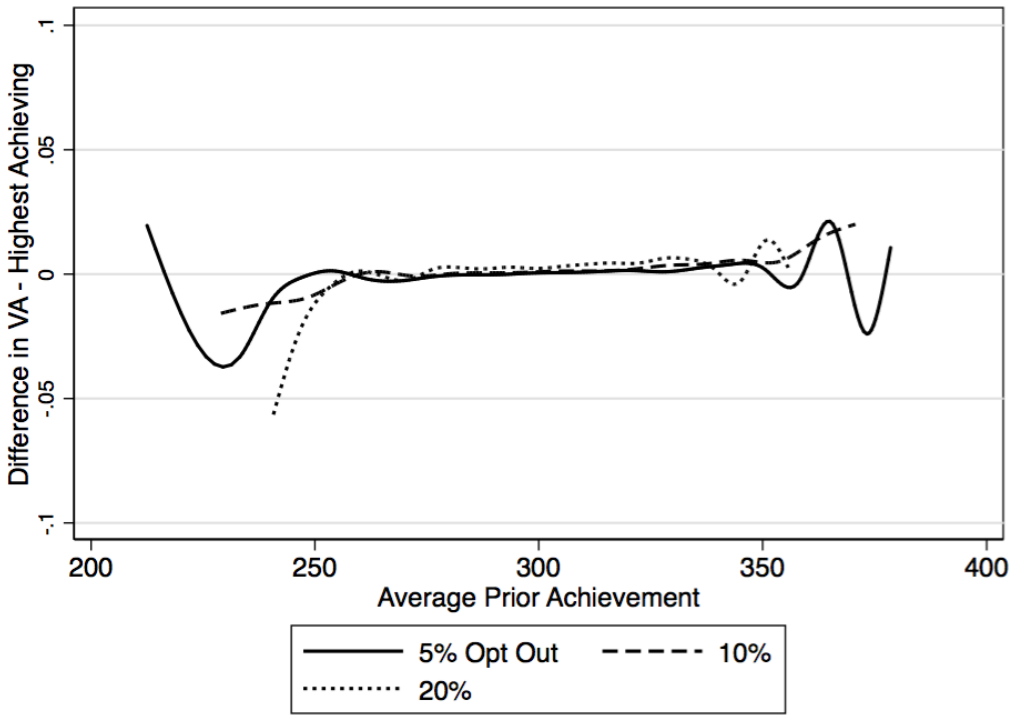


Figure 19: Difference in Complete and Incomplete VA Estimates by Prior Achievement: Lowest Achieving Condition (mspline smoothing, bands = 25)

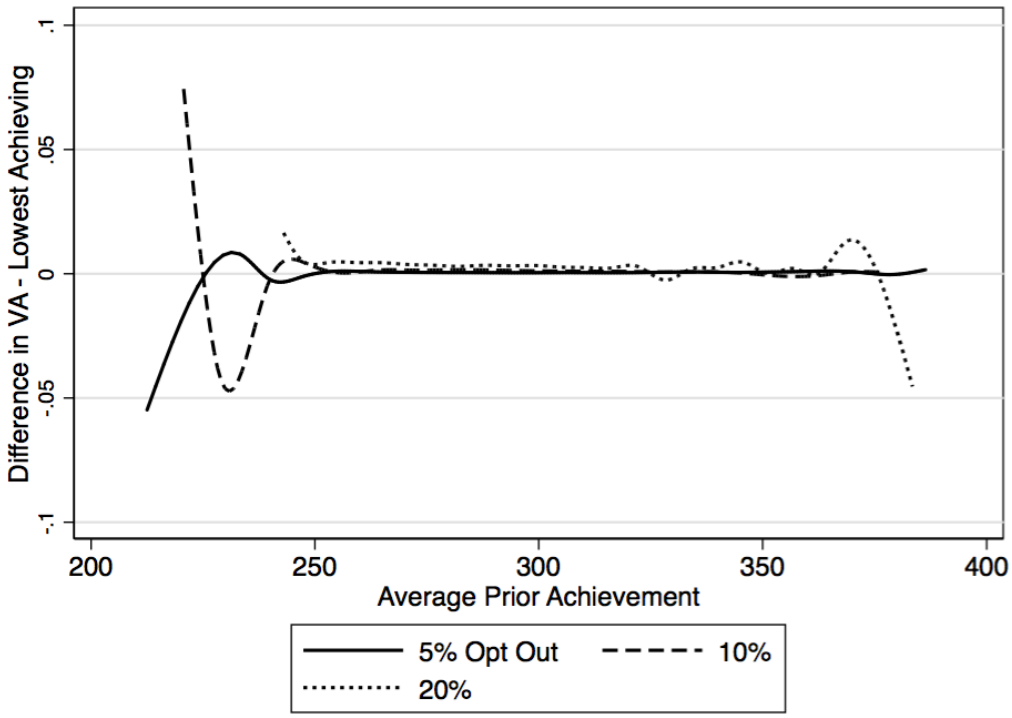


Figure 20: Difference in Complete and Incomplete VA Estimates by Percent Opt Out: Random Condition (mspline smoothing, bands = 25)

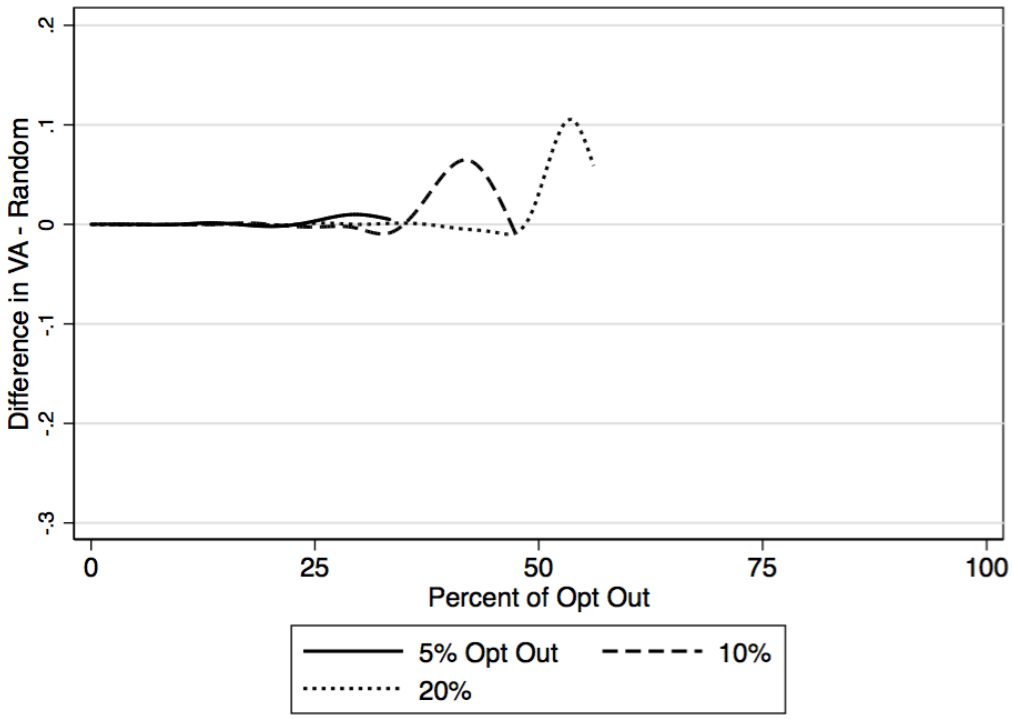


Figure 21: Difference in Complete and Incomplete VA Estimates by Percent Opt Out: Highest Probability Condition (mspline smoothing, bands = 25)

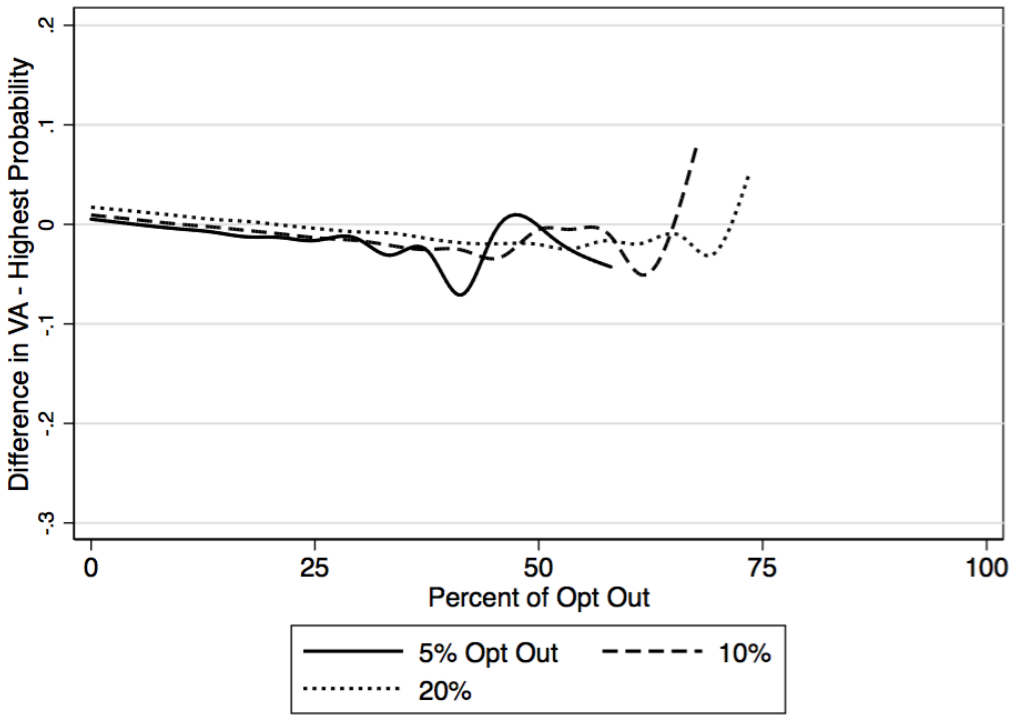


Figure 22: Difference in Complete and Incomplete VA Estimates by Percent Opt Out: Highest Achieving Condition (mspline smoothing, bands = 25)

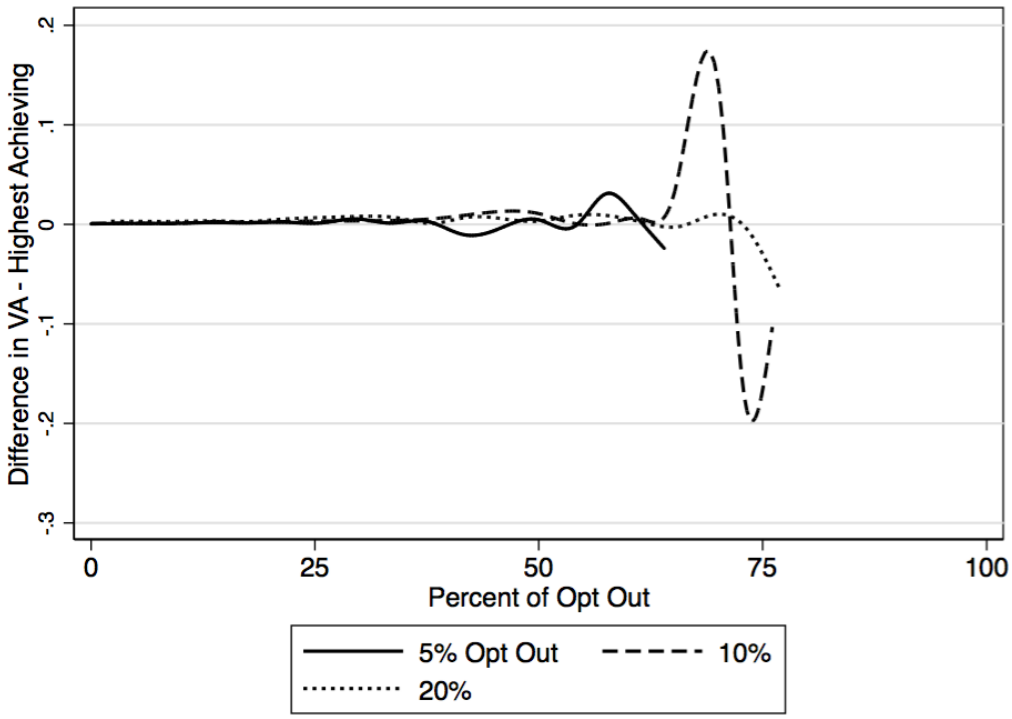


Figure 23: Difference in Complete and Incomplete VA Estimates by Percent Opt out: Lowest Achieving Condition (mspline smoothing, bands = 25)

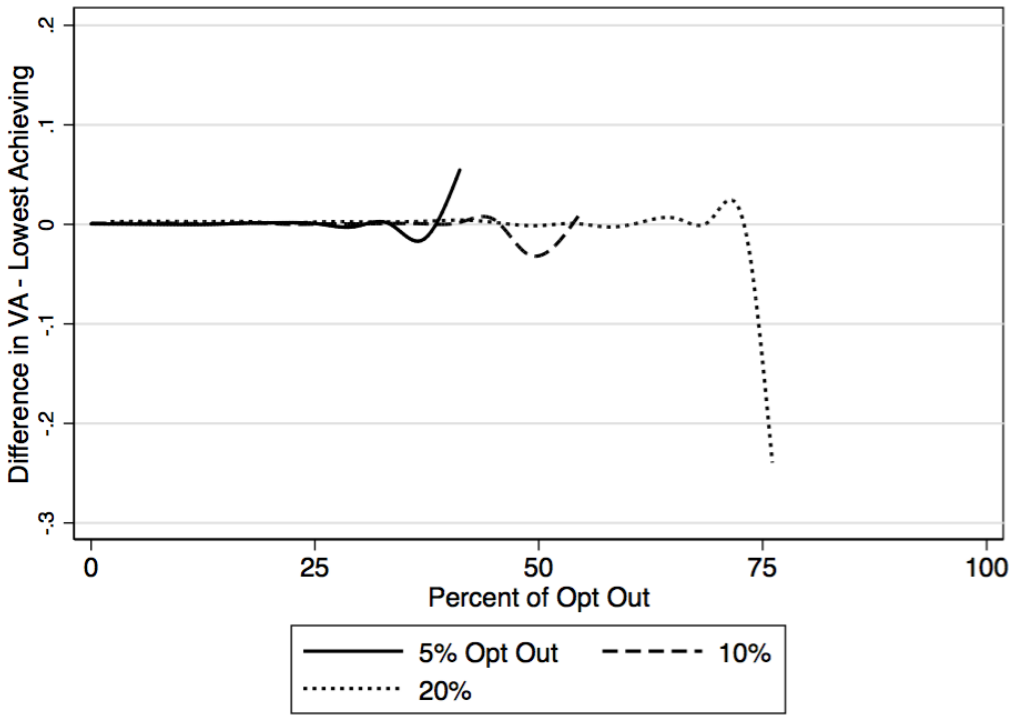


Figure 24: Difference in Complete and Incomplete VA Standard Errors by Prior Achievement: Random Condition (mspline smoothing, bands = 25)

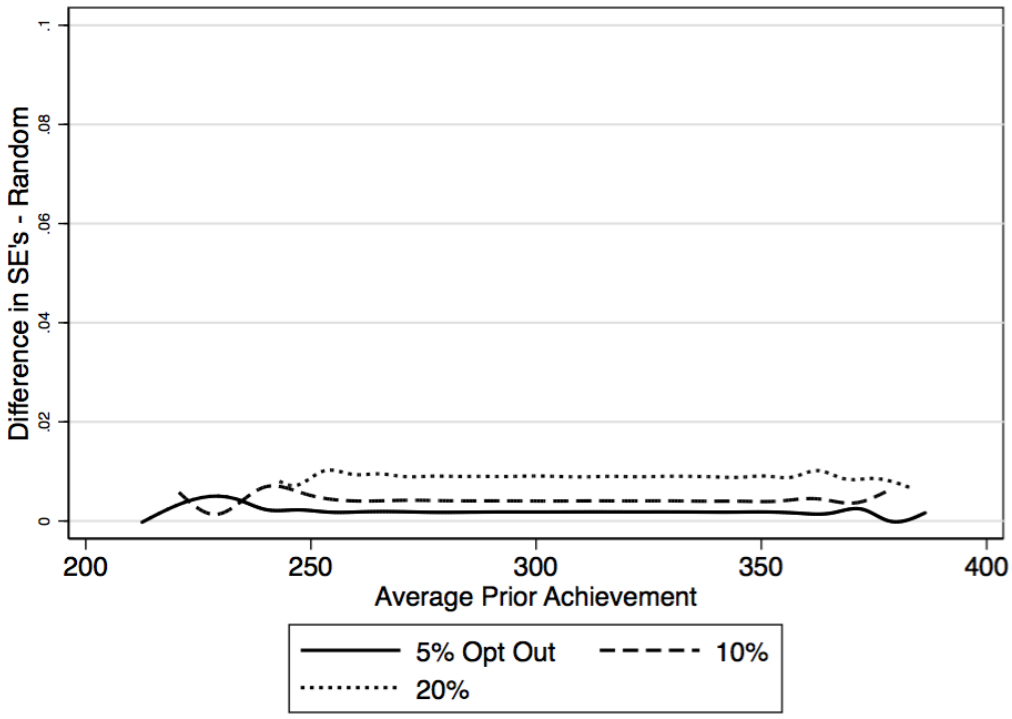


Figure 25: Difference in Complete and Incomplete VA Standard Errors by Prior Achievement: Highest Probability Condition (mspline smoothing, bands = 25)

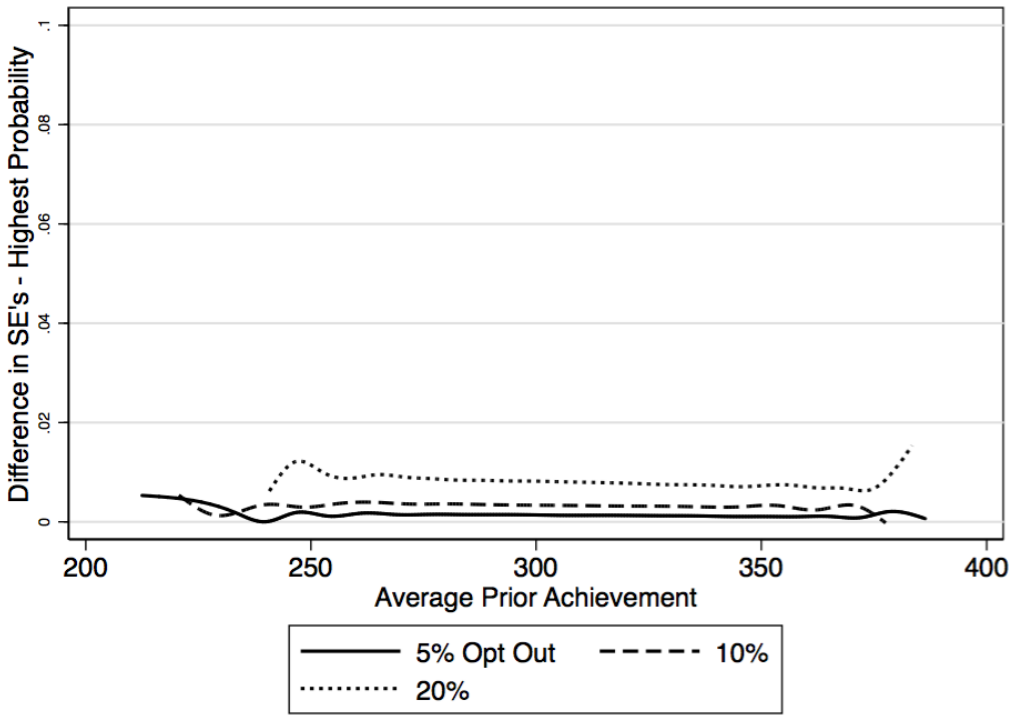


Figure 26: Difference in Complete and Incomplete VA Standard Errors by Prior Achievement: Highest Achieving Condition (mspline smoothing, bands = 25)

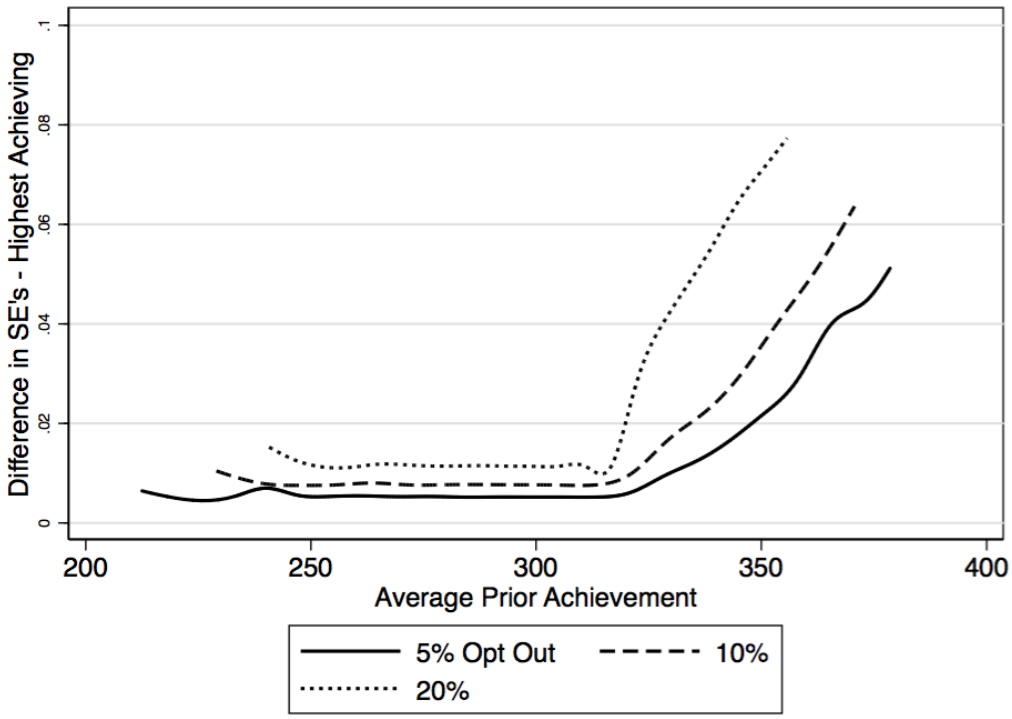


Figure 27: Difference in Complete and Incomplete VA Standard Errors by Prior Achievement: Lowest Achieving Condition (mspline smoothing, bands = 25)

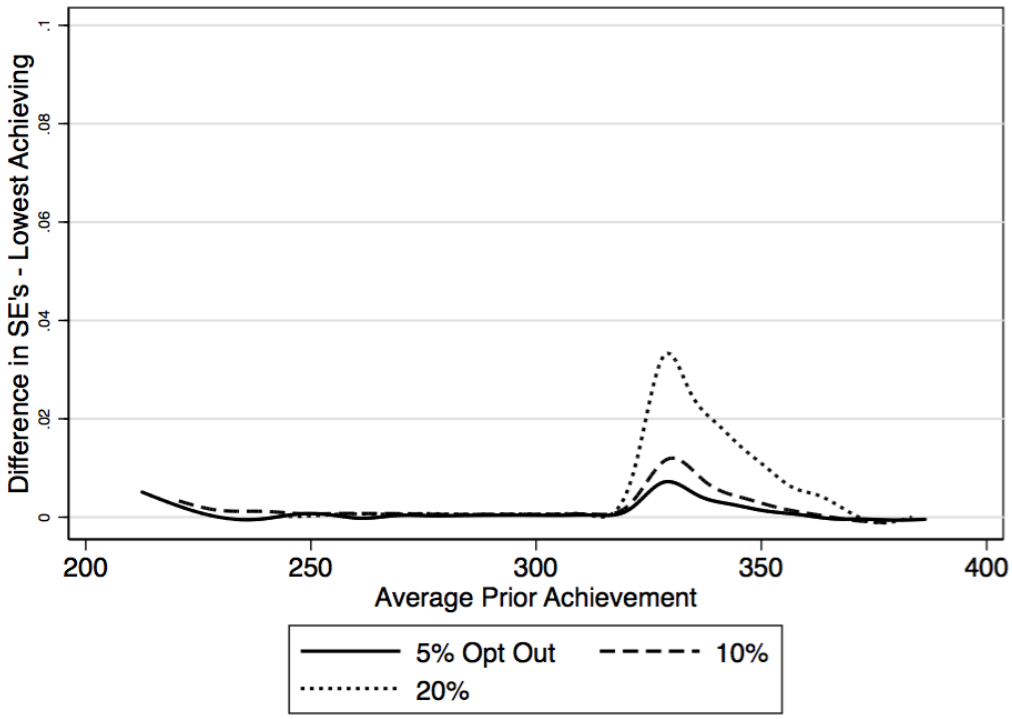


Figure 28: Difference in Complete and Incomplete VA Standard Errors by Percent Opt Out: Random Condition (mspline smoothing, bands = 25)

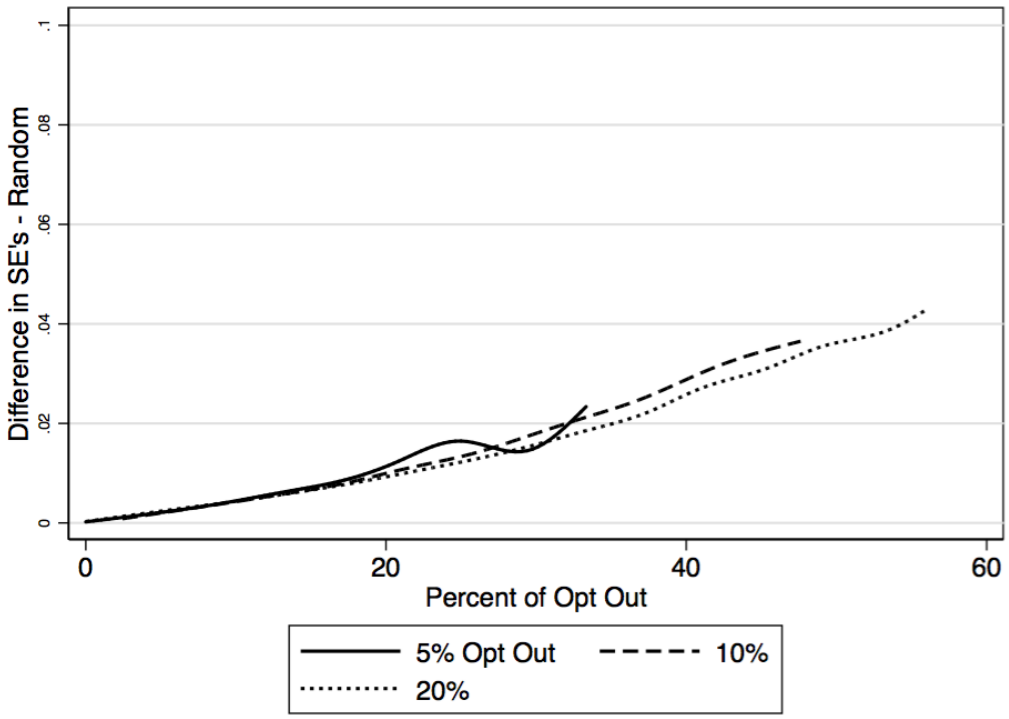


Figure 29: Difference in Complete and Incomplete VA Standard Errors by Percent Opt Out: Highest Probability Condition (mspline smoothing, bands = 25)

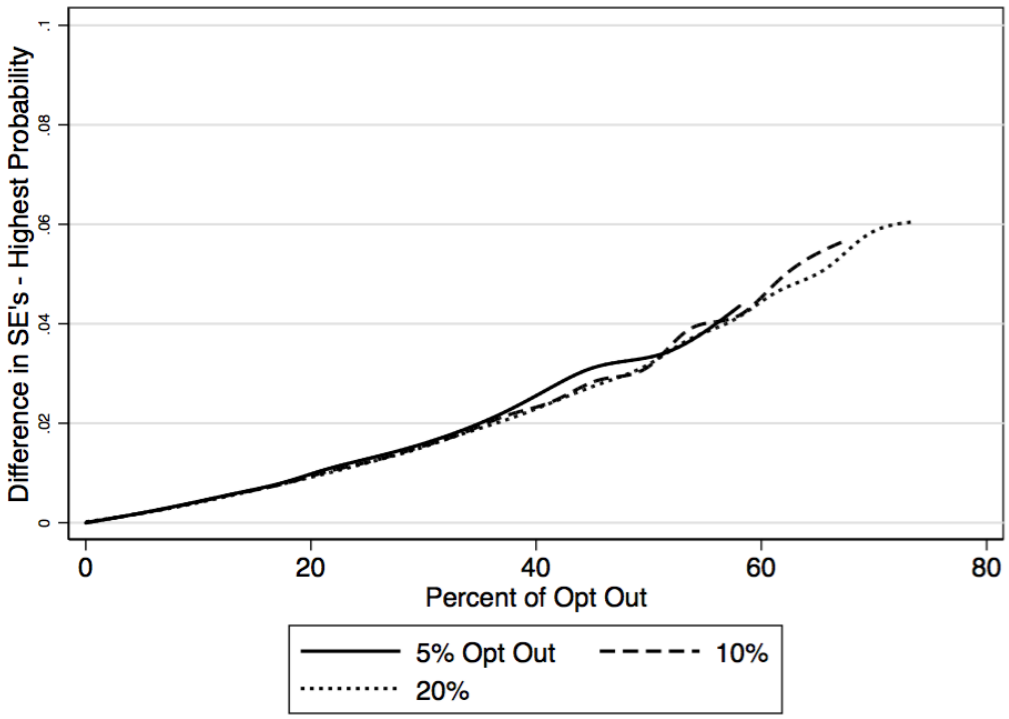


Figure 30: Difference in Complete and Incomplete VA Standard Errors by Percent Opt Out: Highest Achieving Condition (mspline smoothing, bands = 25)

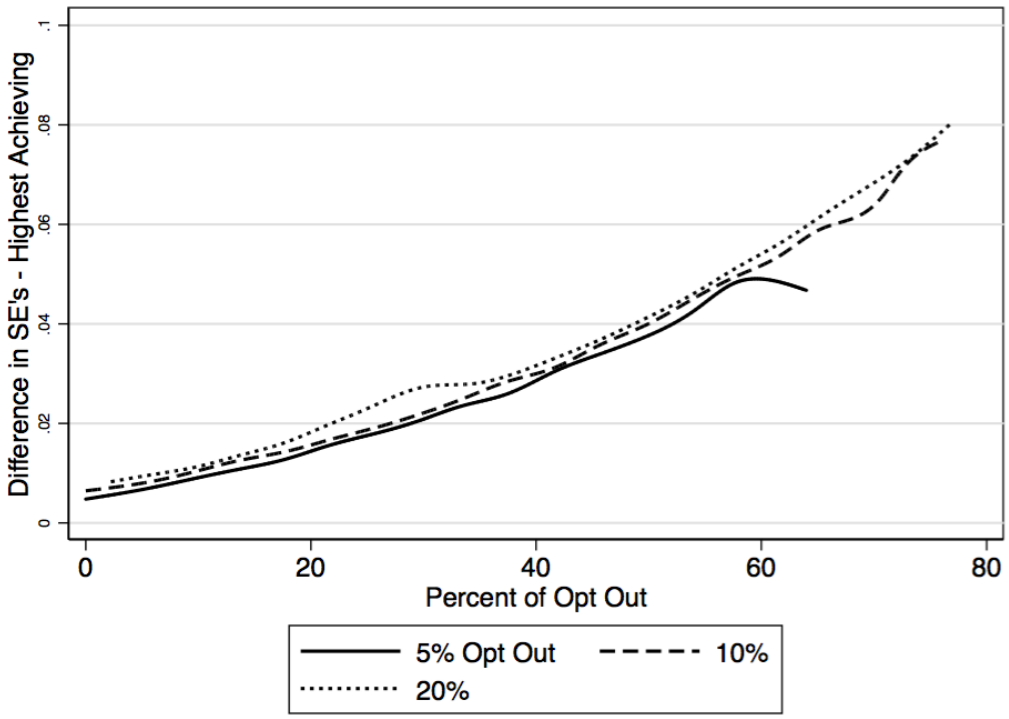


Figure 31: Difference in Complete and Incomplete VA Standard Errors by Percent Opt Out: Lowest Achieving Condition (mspline smoothing, bands = 25)

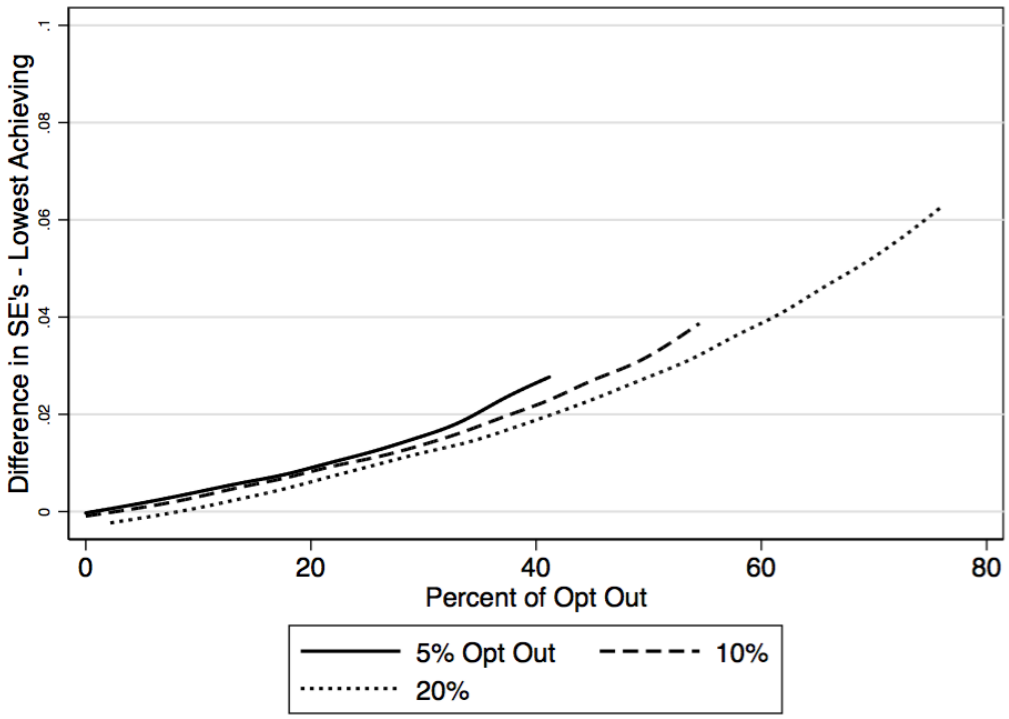


Figure 32: Change in Value-Added Quartile by Complete and Incomplete VA in 20 percent Random Condition

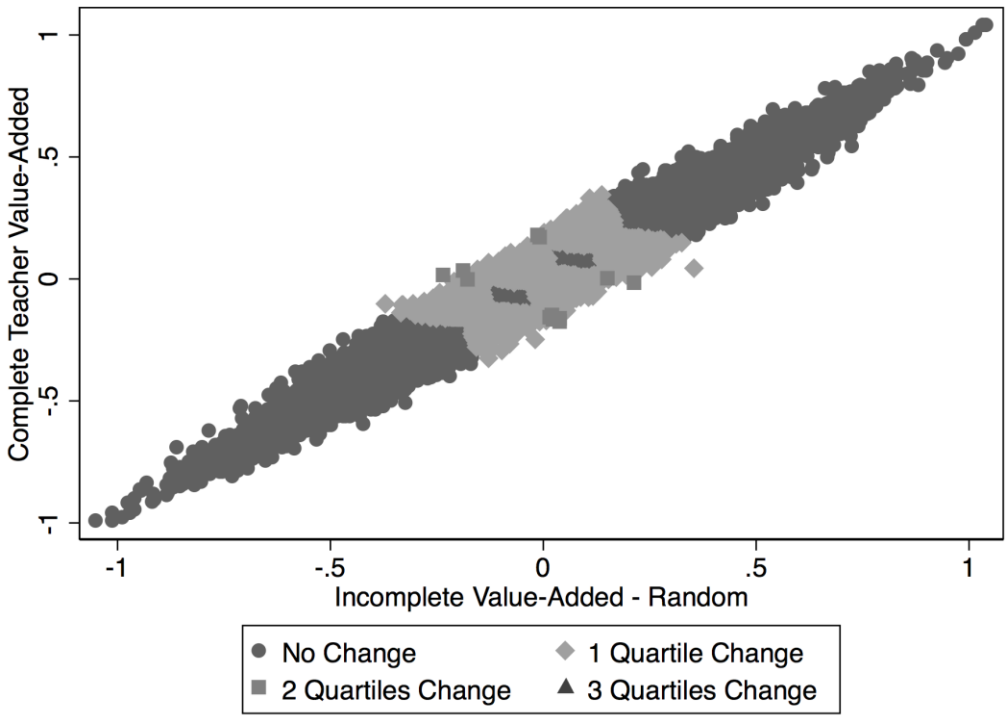


Figure 33: Change in VA Quartile by Complete and Incomplete VA in 20 Percent Highest Probability Condition

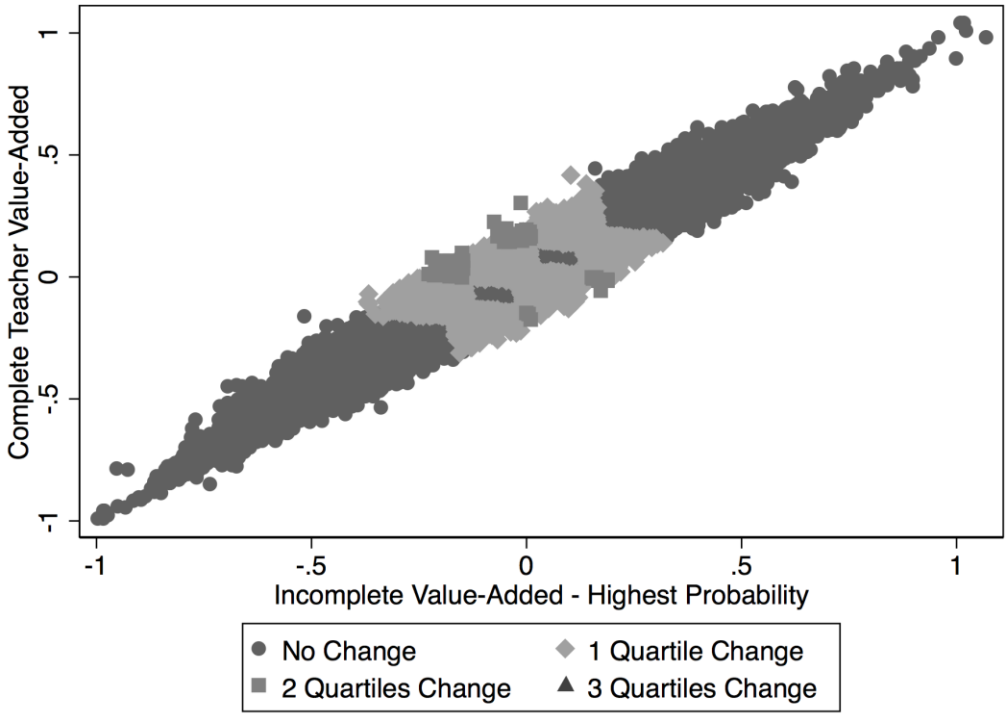


Figure 34: Change in VA Quartile by Complete and Incomplete VA in 20 Percent Highest Achieving Condition

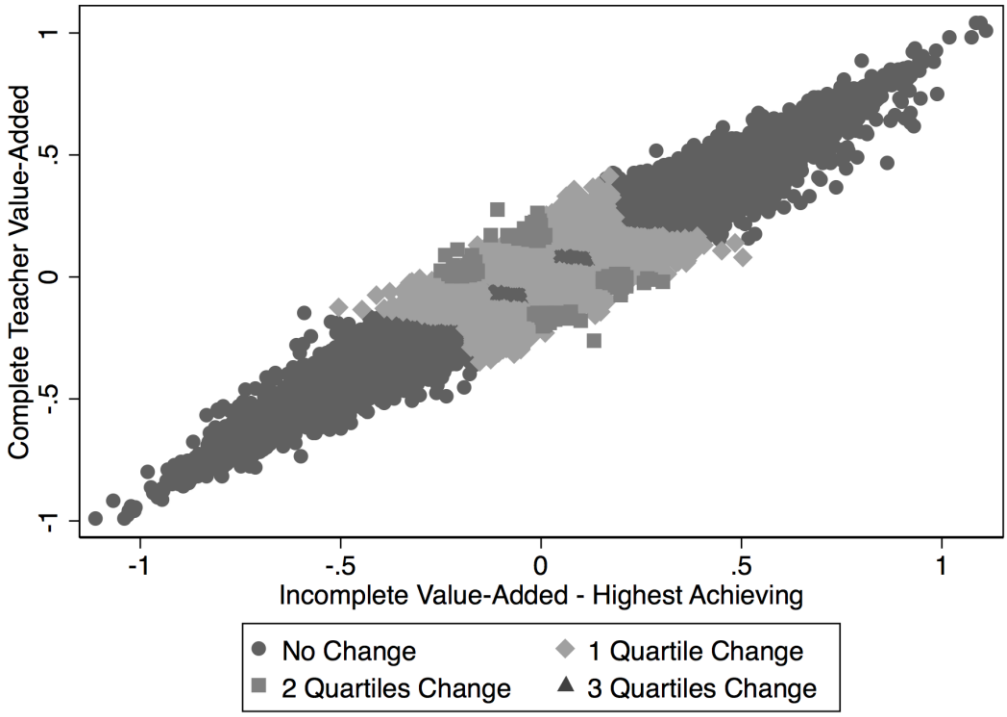


Figure 35: Change in VA Quartile by Complete and Incomplete VA in 20 Percent Highest Achieving Condition

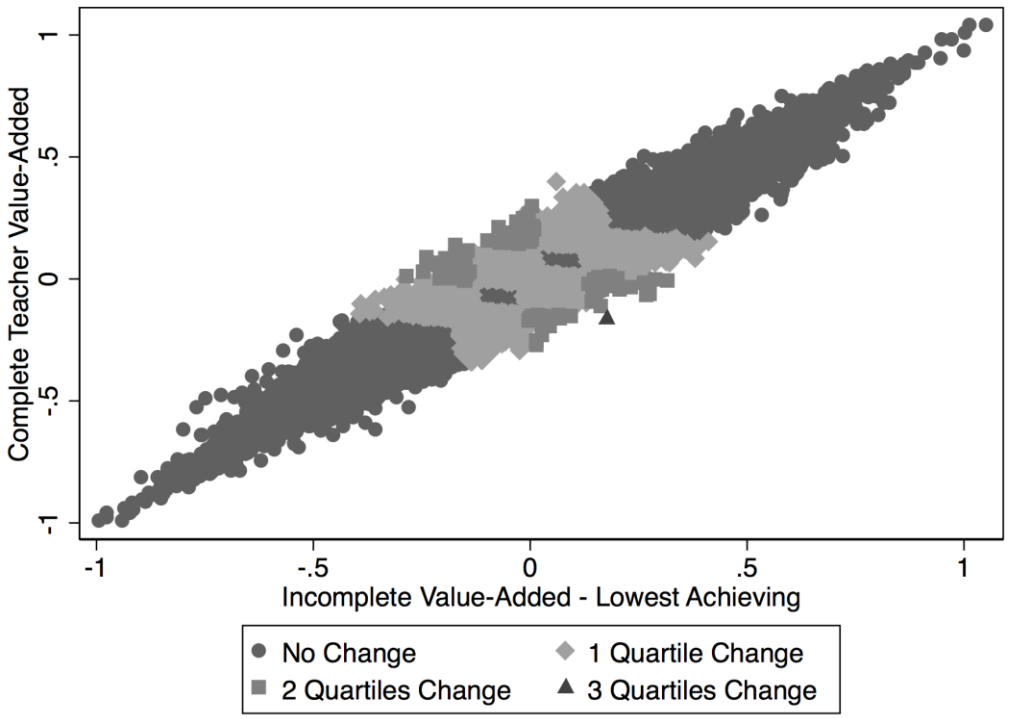


Figure 36: Change in VA Quartile by Complete VA and Average Prior Achievement in 20 Percent Random Condition

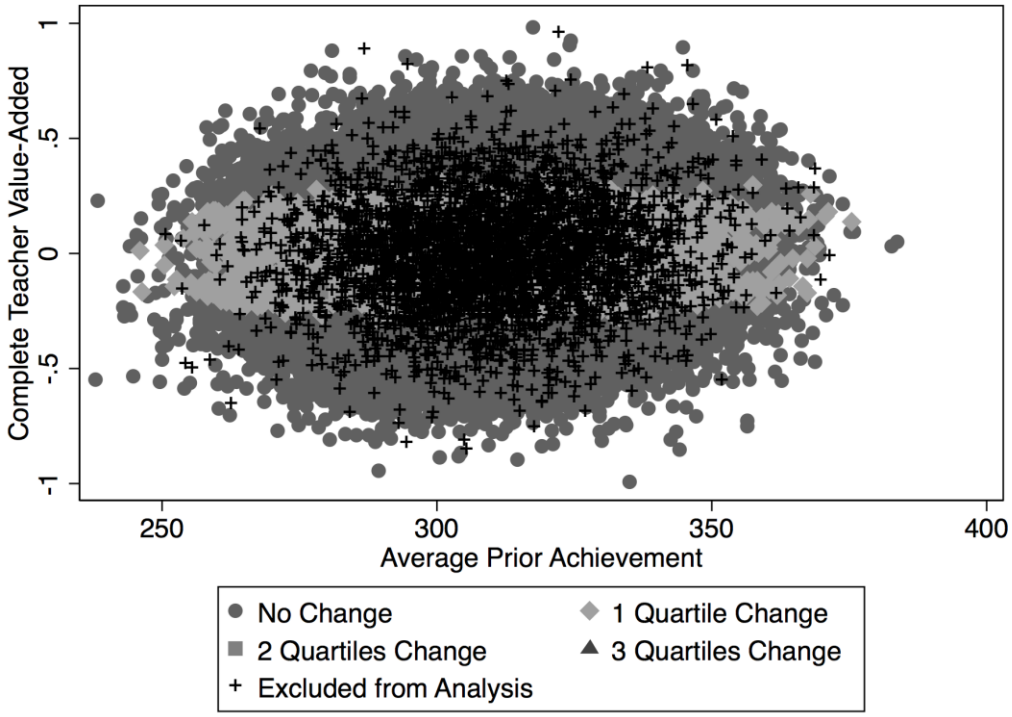


Figure 37: Change in VA Quartile by Complete VA and Average Prior Achievement in 20 Percent Highest Probability Condition

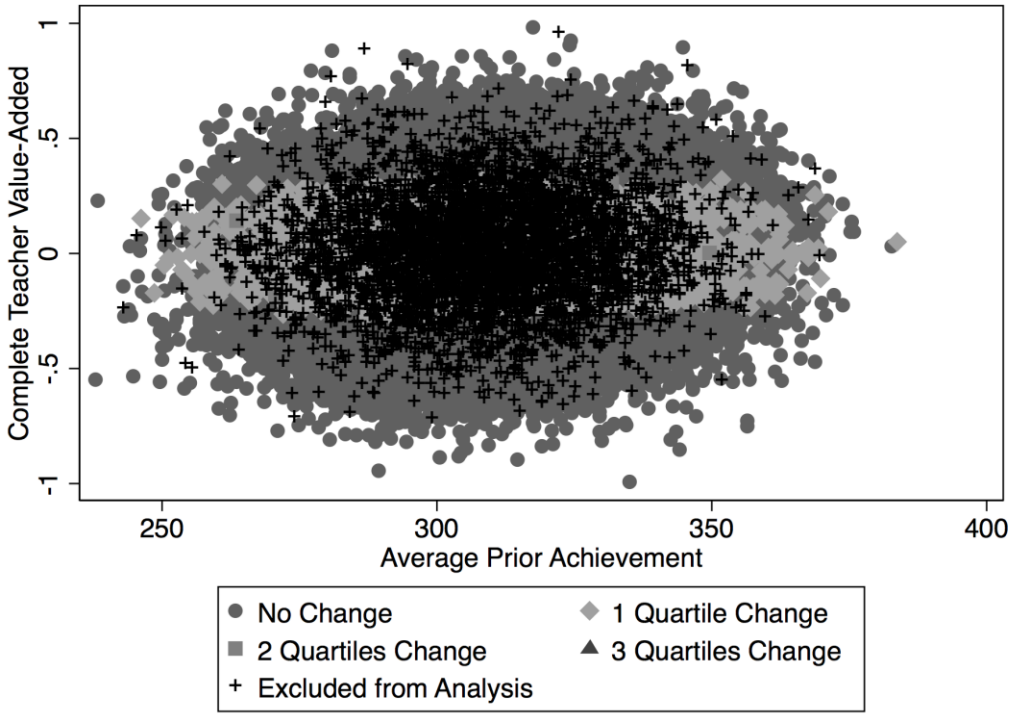


Figure 38: Change in VA Quartile by Complete VA and Average Prior Achievement in 20 Percent Highest Achieving Condition

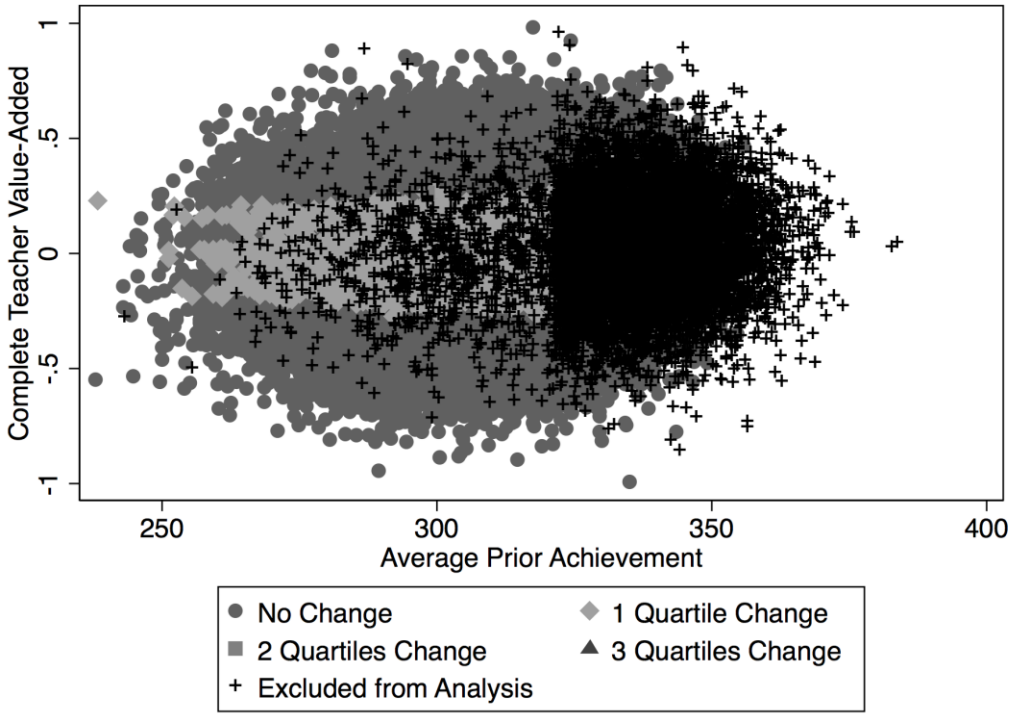
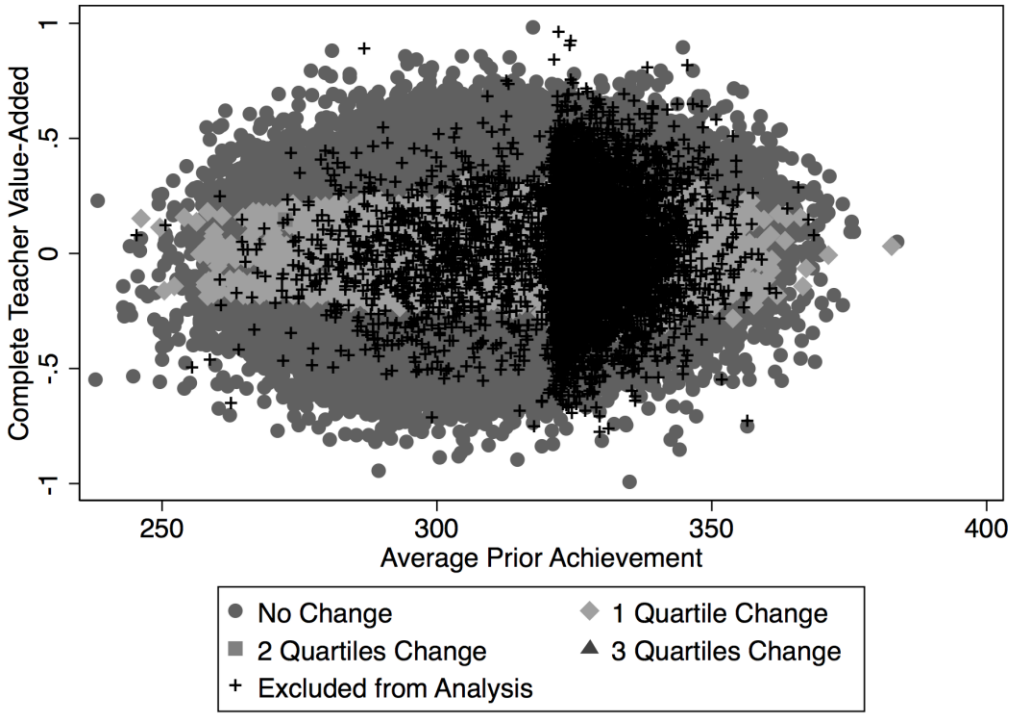


Figure 39: Change in VA Quartile by Complete VA and Average Prior Achievement in 20 Percent Lowest Achieving Condition



REFERENCES

No Child Left Behind (NCLB) Act of 2001, 20 U.S.C.A. § 6301 et seq. (West 2003).

Albright, J. (2016, January 14). Veto override on testing opt-out fails in house. *The News Journal.* Retrieved from: http://www.delawareonline.com/story/news/education/2016/01/14/opt-out-vote/78785656/;

American Institutes for Research. (2016). *2014–15 growth model for educator evaluation: Technical report.* (Technical Report). Washington, DC: American Institutes for Research. Retrieved from https://www.engageny.org/resource/technical-report-growth-measures-2014-15

Bakeman, J. (2015). As Race to the Top ends, controversy continues. *Politico.Com.* Retrieved from http://www.politico.com/states/new-york/albany/story/2015/07/as-race-to-the-top-ends-controversy-continues-023795

Baker, E., Barton, P., Darling-Hammond, L., Haertel, E., Ladd, H., Linn, R., . . . Shepard, L. (2010). *Problems with the use of student test scores to evaluate teachers.* (Research Report No. 278). Washington, DC: Economic Policy Institute.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers *Journal of Educational and Behavioral Statistics, 29*(1), 37-65.

Bennett, R. (2016). *Opt out: An examination of issues.* (ETS Research Report Series). Princeton, NJ: Educational Testing Services. doi: 10.1002/ets2.12101

Blank, R. K. (2010). *State growth models for school accountability: Progress on development and reporting measures of student growth.* (Policy). Washington, D.C.: Council of Chief State School Officers.

Butler, M. A., & Gopalakrishnan, A. (2016). *Improving participation rates on state assessments.* (Policy Report). Hartford, CT: Connecticut Department of Education.

Camera, L. (2015). States seek guidance in face of 'opt out' push. *Education Week, 34*(26), 15-17. Retrieved from http://www.edweek.org/ew/articles/2015/04/01/states-seek-guidance-in-face-of-opt-out.html

Cameto, R., Knokey, A. M., Nagle, K., Sanford, C., & Blackorby, J. (2009). *State profiles on alternate assessments based on alternate achievement standards: A report from the national study on alternate assessments.* (Policy No. NCSER 2009-3013). Washington, D.C.: United States Department of Education. Retrieved from https://ies.ed.gov/ncser/pdf/20093013.pdf

Castellano, K. E., & Ho, A. D. (2013). *A practitioner's guide to growth models.* (Research Report). Washington, DC: Council of Chief State School Officers. Retrieved from http://www.ccsso.org/Resources/Publications/A_Practitioners_Guide_to_Growth_Models.html

Castellano, K. E., & Ho, A. D. (2014). *Practical differences among aggregate-level conditional status metrics: From median student growth percentiles to value-added models.* (Technical Report). Cambridge, MA: Harvard University. Retrieved from http://scholar.harvard.edu/andrewho/publications/practical-differences-among-aggregate-level-conditional-status-metrics-median

Chingos, M. (2014). *Who opts out of state tests?* (Research Report). Washington, DC: Brookings Institution. Retrieved from http://www.brookings.edu/research/papers/2015/06/18-chalkboard-who-opts-out-chingos

Chism, M. (2015). In California Department of Education (Ed.), *Participation rate letter to the state of California*. Washington, D.C.: United States Department of Education.

Clark, A. (2015, January 30). Assembly bill would allow students to opt out of testing without penalty, sponsor says. *NJ.Com* Retrieved from http://www.nj.com/politics/index.ssf/2015/01/parcc_opt-out_bill_introduced_in_nj_assembly.html

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review, 26*(6), 673-682.

Colorado Department of Education. (2016). 2016 state assessment score release - participation stabilizes, improvement seen but more work to be done. Retrieved from http://www.cde.state.co.us/communications/20160811assessmentstatelevelscorerelease

Dieterle, S., Guarino, C., Reckase, M., & Wooldridge, J. (2012). *How do principals assign students to teachers: Finding evidence in administrative data and the implications for value-added.* (Working Paper No. 30). Lansing, MI: Michigan State University Education Policy Center.

Elledge, A., Le Floch, K. C., Taylor, J., & Anderson, L. (2009). *State and local implementation of the "No Child Left Behind act". Volume V--implementation of the 1 percent rule and 2 percent interim policy options.* (Research Report). Washington, DC: U.S. Department of Education.

Guarino, M., Reckase, M., Stacy, B., & Wooldridge, J. (2015). *A comparison of growth percentile and value-added measures of teacher performance.* (Working No. 39). Lansing, MI: The Education Policy Center at Michigan State University.

Hammond, B. (2015, June 9). Oregon risks losing $140 million for enabling kids to skip common core tests, feds warn. *The Oregonian* Retrieved from http://www.oregonlive.com/education/index.ssf/2015/06/new_oregon_testing_law_could_j.html

Harris, E. (2015). 20% of New York state students opted out of standardized tests this year. *New York Times* Retrieved from http://www.nytimes.com/2015/08/13/nyregion/new-york-state-students-standardized-tests.html?_r=0

Harris, E. A., & Fessenden, F. (2015). 'Opt out' becomes anti-test rallying cry in new york state. *New York Times* Retrieved from http://nyti.ms/1c4ZMBp

Henderson, M. B., Peterson, P. E., & West, M. (2015). *The 2015 EdNext poll on school reform.* (No. 16.1). Cambridge, MA: Education Next. Retrieved from http://educationnext.org/2015-ednext-poll-school-reform-opt-out-common-core-unions/

Institute of Education Sciences. (2007). *National assessment of Title I final report: Summary of key findings.* (No. NCEE 2007-4014). Washington, D.C.: United States Department of Education.

Kalogrides, D., Loeb, S., & Beteille, T. (2013). Systematic sorting: Teacher characteristics and class assignments. *Sociology of Education, 86*(2)

Kane, T., & Staiger, D. (2010). *Are value-added measures of teaching effectiveness too volatile to use?* (White Paper). Cambridge, MA: Center for Education Policy Research.

Karl, A. T., Yang, Y., & Lohr, S. (2013). A correlated random effects model for nonignorable missing data in value-added assessment of teacher effects. *Journal of Educational and Behavioral Statistics, 38*(6), 577–603. doi:10.3102/1076998613494819

Katz, S. (2013). *Information on student participation in state assessments.* Unpublished manuscript. Retrieved March 1, 2016, Retrieved from http://www.wpcsd.org/downloads/student-participation.pdf

Klein, A. (2015, December 22). Ed. dept. to states: Even under ESSA, you need a plan for high opt-out rates. Retrieved from http://blogs.edweek.org/edweek/campaign-k-12/2015/12/ed_dept_to_states_under_essa_need_plan_for_opt-Outs.html

Lorenzo, S. J. (2015). *Opt-out policies by state.* (Policy Report No. 6). Alexandria, VA: National Association of School Boards of Education.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2004). *Evaluating value-added models for teacher accountability.* (Research Report No. MG-158). Santa Monica, CA: RAND Corporation.

McCaffrey, D., Lockwood, J. R., Mariano, L. T., & Setodji, C. (2005). Challenges for value-added assessment of teacher effects. In R. Lissitz (Ed.), *Value-added models in education: Theory and applications* (pp. 111-144). Maple Grove, MN: JAM Press.

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

McLaughlin, D., Gallagher, L., & Stancavage, F. (2004). *Evaluation of bias correction methods for "worst-case" selective non-participation in NAEP*. (Technical Report). Washington, DC: American Institutes for Research.

Meyer, R. (1994). *Educational performance indicators: A critique.* (Discussion Paper No. 1052-94). Madison, WI: Institute for Research on Poverty, University of Wisconsin-Madison.

National Center on Educational Outcomes, Council of Chief State School Officers, & National Association of State Directors of Special Education. (2005). Uneven transparency: NCLB tests take precedence in public assessment reporting for students with disabilities. (Technical Report 43). Minneapolis, MN: National Center on Educational Outcomes, University of Minnesota.

New Jersey Department of Education. (2015). Meeting participation targets for New Jersey State assessments: Action plan development guide. Retrieved from http://www.state.nj.us/education/title1/accountability/progress/15/ActionPlan.pdf

New York State Education Department. (2015). *State education department releases spring 2015 grades 3-8 assessment results.* (Press Release). Albany, NY: New York State Education Department. Retrieved from http://www.nysed.gov/news/2015/state-education-department-releases-spring-2015-grades-3-8-assessment-results#_ftn2

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives, 28*(3) Retrieved from http://epaa.asu.edu/ojs/article/view/810

Papay, J. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163-193. doi:10.3102/0002831210362589

Parr, A., & Teed, P. (2015). *2014-15 assessment results.* (Policy Memo). Olympia, WA: Washington State Board of Education.

Phi Delta Kappan, & Gallup. (2015). The 47th PDK/Gallup poll of the public's attitudes toward the public schools: Testing doesn't measure up for Americans. *Phi Delta Kappan*, 27(1)

Piesse, A., Rust, K., & National Center for, E. S. (2003). *U.S. 2001 PIRLS nonresponse bias analysis.* (*Working Paper No. 2003-21*). Washington, DC: National Center for Education Statistics.

Pizmony-Levy, O., & Green Saraisky, N. (2016). *Who opts out and why? results from a national survey on opting out of standardized tests.* (Research Report). New York: Teachers College, Columbia University.

Redfield, D., & Sheinker, J. (2004). *Framework for transitioning from IASA to NCLB.* (Policy Report). Washington, D.C.: Council of Chief State School Officers.

Reid, K. S. (2014). Testing skeptics' advice: Just say 'no'. *Education Week, 33*(24), 1-19. Retrieved from http://www.edweek.org/ew/articles/2014/03/12/24boycotts_ep.h33.html

Rice, A., Marland, J., & Meyer, R. (2016). The impact of student assessment opt out on achievement and growth metrics in New York state. *Paper Presented at the Association for Educational Finance and Policy Annual Conference,* Denver, CO.

Rockoff, J., Staiger, D., Kane, T., & Taylor, E. (2010). *Information and employee evaluation: Evidence from a randomized intervention in public schools.* (Working Paper No. 16240). Washington, DC: National Bureau of Economic Research.

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

Rowland-Woods, J., Wixom, M., & Aragon, S. (2015). *Assessment opt-out policies: State responses to parent pushback.* (Policy Report). Washington, DC: Education Commission of the States. Retrieved from http://www.ecs.org/assessment-opt-out-policies-state-responses-to-parent-pushback/

Rubin, D. (1987). In Barnett R., Bradley R., Hunter J. S., Kendall D., Miller R., Smith A., . . . Watson G. (Eds.), *Multiple imputation for nonresponse in surveys.* (1st ed.). New York, NY: Wiley & Sons.

Sass, T. (2008). *The stability of value-added measures of teacher quality and implications for teacher compensation policy.* (Research Brief No. 4). Washington, DC: The Urban Institute.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177. doi:10.1037/1082-989X.7.2.147

The National Center for Fair and Open Testing. (2015). More than 620,000 refused tests in 2015. Retrieved from http://www.fairtest.org/more-500000-refused-tests-2015

Ujifusa, A. (2015, June 23). Testing opt-out bill signed by Oregon Gov. Kate Brown; Delaware next? Retrieved from http://blogs.edweek.org/edweek/state_edwatch/2015/06/testing_opt-out_bill_signed_by_oregon_gov_kate_brown_delaware_next.html

Ujifusa, A. (2015, December 23). Education department asks 13 states to address low test-participation rates. Retrieved from http://blogs.edweek.org/edweek/campaign-k-12/2015/12/twelve_states_asked_to_address.html

United States Department of Education. (2011). Why participation in the assessment is important. Retrieved from https://nces.ed.gov/nationsreportcard/about/natimportant.aspx

United States Department of Education. (2013). *State and local report cards*. Washington, D.C.: United States Department of Education. Retrieved from http://www2.ed.gov/programs/titleiparta/state_local_report_card_guidance_2-08-2013.pdf

United States Department of Education. (2016a). Race to the Top fund: Score of work decision letters. Retrieved from http://www2.ed.gov/programs/racetothetop/awards.html

United States Department of Education. (2016b). ESEA flexibility: State requests and related documents. Retrieved from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html

Walker, B. (2015, August 20). New Jersey senate passes PARCC opt-out resolution. The Heartland Institute Retrieved from http://news.heartland.org/newspaper-article/2015/08/20/new-jersey-senate-passes-parcc-opt-out-resolution