A mixed-methods investigation of how schools in an urban district respond to student benchmark data under shifting accountability incentives

By

Emily C. Kern

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Leadership and Policy Studies

August 9, 2019

Nashville, Tennessee

Approved:

Dale Ballou, Ph.D.

Jason A. Grissom, Ph.D.

William R. Doyle, Ph.D.

Deborah W. Rowe, Ph.D.

To my continuously supportive and amazing husband, Scott, who pushes me to dream bigger

and

To Sergio, whose seventh-grade year inspired this project and who reminds me that policy directly affects students’ lives

ACKNOWLEDGEMENTS

This work would not have been possible without the financial support of the Peabody graduate honor scholarship or the Bonsal Applied Education Research Award.

I am supremely grateful for the privilege to grow professionally and to learn continuously from the unbelievably accomplished researchers who make up my dissertation committee. To Dr. Dale Ballou, who patiently read multiple iterations of this project and whose feedback has strengthened this work and made me a better researcher. To Dr. Jason Grissom, whose encouragement, guidance, and mentorship supported me throughout the program. To Dr. Will Doyle, who helped move me from never having written a line of code to a confident quantitative researcher. To Dr. Matthew Springer, who offered insight into triage work and the opportunity to teach at the college level. And to Dr. Deborah Rowe, who trained me in qualitative methods and who encouraged this project's focus on what is happening in classrooms.

I am grateful to all of those with whom I have had the pleasure to work in a variety of other research settings. To Dr. Paul Cobb, whose research-practice partnership work that pushes to increase students' mathematical conceptual understanding represents the type of research that I benefitted from as a teacher and a coach and who provided numerous research opportunities. To Dr. Erin Henrick, whose contagious optimism and unrelenting interest in improving instruction made working together a joy. To the rest of the MIST, MIST2, and PMRR crew for the countless hours spent talking about my first love: quality math teaching.

I could not have made it through these past years without the study sessions and encouragement from my cohort (Samantha, Seth, Adam, Chris, and Luis) or the stress relief provided by the Nashville Rollergirls.

To the numerous teachers, coaches, mentors, administrators, and district leaders I have worked with over the past fifteen years who keep the focus on student learning.

To my inspirations for life-long learning: my dad David, and Julie Amaon.

Nobody has been more important to me in the pursuit of this project than the members of my family. I am a third-generation teacher inspired by a long line of strong women: my mother Ann, her mother Sara, and her mother Lottie. I would like to thank my parents, whose love and unconditional support have always allowed me to choose my own path. Thank you to my three brothers who helped me develop a strong sense of fairness. Most importantly, I wish to acknowledge and appreciate the support from my loving and ever-patient husband, Scott.

LIST OF TABLES

LIST OF FIGURES

Figure Page

TABLE OF CONTENTS

Appendix

Chapter 1: Introduction

Congress passed No Child Left Behind (NCLB) nearly two decades ago, which expanded the federal government's role in education and held schools accountable for student learning by directly tying federal funds to student test scores (Manna, 2011). Although NCLB's metrics were relatively prescribed, the federal government has since given states increasingly more freedom to design the incentives of their school accountability policies. Legislators passed NCLB with the dual purposes of (a) raising achievement for all students and (b) closing the achievement gap (McGuinn, 2006). The policymakers' equity intentions are illustrated by the name of the policy. These intentions, however, did not necessarily align with the metrics used to rate schools. How the state determined school ratings mattered because NCLB was the first federal accountability policy where schools faced consequences for failing to meet the annual measureable objectives set by the state (McGuinn, 2006). These sanctions escalated each subsequent year that a school missed accountability targets, ranging from allowing students to transfer schools to restructuring continuously-failing schools (which included the threat of state take-over and job loss) (Manna, 2011). NCLB's consequences followed the theory of accountability, that the combination of measurement, transparency (school ratings were released publicly), and consequences would motivate schools to improve (Manna, 2011).

NCLB's consequences, coupled with the metrics used to evaluate schools, created perverse incentives for schools to meet rising student achievement requirements. Student achievement was measured by the *proficiency rate*, or the percentage of students in a school who met proficiency on the state test. All-or-nothing proficiency labels do not account for student starting points because students either pass or do not, regardless of how much improvement they demonstrate over the previous year. Rating schools on this metric incentivizes them to focus on students who are close to meeting proficiency (sometimes called "bubble" students) at the expense of low- and high-performers. Diverting resources away from students who are so low that they are unlikely to meet proficiency and away from those who are likely pass and towards students close to proficiency is a strategic behavior known as *educational triage* (Gillborn & Youdell, 1999). This unintended consequence of NCLB means that the students with the most need, those who score the lowest on the achievement tests, may be neglected by schools as they focus on bubble students.

In response to growing criticism of NCLB and Congress' failure to reauthorize the policy, the Department of Education in 2011 allowed states to apply for waivers from NCLB's requirements by proposing an alternative accountability system (Polikoff, McEachin, Wrabel, & Duque, 2014). These waivers allowed states to adjust the indicators used to hold schools accountable for student learning. By March 2013, 48 states submitted requests for waivers (Polikoff et al., 2014). Although schools continued to be evaluated on proficiency rates, states implemented a variety of other incentive changes, and many states added growth metrics to their revised accountability systems (Riddle, 2012). A growth system of accountability would theoretically incentivize schools to focus more attention on lower-performing students because, even if these students cannot improve enough in a given year to get above the proficiency line, they can still make gains.

Despite concerns that proficiency rates cause schools to neglect low-performing students, empirical evidence of educational triage is mixed. Qualitative and case studies clearly demonstrate that triage happens (e.g., Booher-Jennings, 2005; Diamond & Spillane, 2004), yet quantitative researchers using large administrative datasets have found mixed evidence regarding this practice (e.g., Ballou & Springer, 2016; Krieg, 2008; Ladd & Lauen, 2010; Reback, 2008;

Springer, 2008b). These mixed findings may result from limitations of administrative datasets that often lack (a) clear indicators of which students schools would consider “on the bubble” of proficiency and (b) measures of resource allocation within the school.

Some of these limitations are addressed in this project by using a unique dataset from a large urban district. Between 2009-10 and 2013-14, Greenfield County Public Schools (GCPS, a pseudonym) implemented a benchmark test policy. Like many other large urban districts (Burch, 2010), GCPS adopted the benchmark policy in response to the high-stakes testing of NCLB. This initiative included (a) assessing third through eighth grade students three times per year in math and reading, (b) projecting student end-of-year performance on the state test, and (c) sharing that information with schools. The district provided data coaches to support administrators and teachers in using this benchmark data to inform instructional decisions. During these five years, the state changed from NCLB to a waiver accountability system. The waiver introduced a suite of changes to the accountability system, which included adding growth measures to the school accountability and educator evaluation systems. This created a natural experiment that allows for exploring whether the change in policy incentives influenced how schools used the benchmark data shared by the district.

Most quantitative research investigating educational triage relies on assumptions regarding which students would be considered bubble by their schools. If schools in GCPS were going to triage, they might utilize the district-supplied benchmark information. GCPS’ dataset offers the opportunity to test the hypothesis that schools use the benchmark data but not necessarily as intended. A feature of the benchmark information shared with schools is that students were labeled Advanced, Proficient, Basic, or Below Basic on each benchmark, a projection which was based on the number of questions they answered correctly. Although these test-score labels provide no information beyond students' underlying scores, schools may have used them as shortcuts to determine which students should receive additional support (i.e., to identify the bubble students). In addition to sharing these performance labels, during the first year of the waiver accountability system, the district combined multiple metrics to assign students even more prescriptive labels—Multi-Year Plan, Priority, or Enrichment—and encouraged schools to target the Priority students in the two months before the state test. The GCPS benchmark data is utilized to specifically investigate whether low-performing students were harmed during NCLB and, if so, whether the policy changes caused schools to treat lower-performing students more equitably. This project examines how the effect of the test-score labels on student outcomes changes under shifting accountability incentives.

In addition to labeling students based on their benchmark scores, in 2012-13, the district gave 29 schools additional funds to target their Priority students in advance of the state test. These school leaders turned in proposals detailing (a) which students were targeted with interventions and (b) what resources they would provide to those students. These documents describe how schools would restructure their school days to offer targeted interventions. The school proposals are qualitatively analyzed to investigate how school leaders proposed to spend those funds. These documents provide information on school-level resource allocation that is not generally available in administrative datasets.

This dissertation uses mixed methods to explore the effects of accountability policy changes on student outcomes. It takes a systems view that the state accountability system created a set of incentives, which led the district to take a series of actions (e.g., implementing the benchmark policy, providing funding for schools to focus on certain students), which influenced how principals and teachers viewed and treated students. This project investigates what happens at the intersection of state and district policy and explores broad questions regarding the effects of policy incentives on students. Did schools engage in triage during NCLB? If so, when during the school year did that begin? Do changes in policy incentives change how schools behave? Which students gain or lose from which policy incentives? Are the equity goals of federal accountability policy realized? These policy- and research-relevant questions are examined using local benchmark assessment information which offers the most up-to-date information schools had on student learning at different points in the school year. The benchmark data offers insight into how schools use this information, which is useful for district leaders who want to help schools use data for instructional purposes. The proposals provide a unique view inside the black box to explore what strategic behavior looks like. Which students were targeted? What resources were used, and how were interventions structured? The proposals completed by school leaders offer resource allocation information that does not rely on patterns of test score gains, which is beneficial for broadening the discussions of policymakers, district leaders, and researchers.

The rest of this dissertation is arranged as follows. This chapter provides the (a) theoretical and empirical background of educational triage and the effect of performance labels on student outcomes, (b) the state accountability context describing the shifting incentives between 2009-10 and 2013-14, and (c) the GCPS district context describing the benchmark policy. Chapter 2 is a quantitative investigation of the effect of the Below Basic, Basic, Proficient, and Advanced labels over the five years of the benchmark policy. It explores the hypotheses that schools used these test-score labels to triage during NCLB and that this behavior changed after the waiver. Chapter 3 examines alternative explanations for the previous chapter's findings that low-performers gained during the waiver. These alternatives include (a) the adoption of a new educator evaluation system in 2011-12, (b) the more prescriptive district-supplied labels of Multi-Year Plan, Priority, and Enrichment, and (c) the additional funding provided for the select group of schools in 2012-13. Chapter 4 uses qualitative methods to analyze the 29 school documents that detail schools' proposed academic interventions. This chapter investigates which students receive what resources in the schools that were provided additional funding to target Priority students. The dissertation concludes with Chapter 5, which discusses the results from the analyses; the implications for policymakers, district and school leaders, and researchers; and next steps for further research.

Theoretical and Empirical Background

Theoretical Background: Educational Triage

One main difference between NCLB and prior federal education policies was the threat of increasingly consequential sanctions for schools that did not meet yearly benchmarks (McGuinn, 2006). Faced with high-stakes exams in math and reading, schools responded to NCLB's incentives in a variety of unintended ways. For example, schools engaged in a variety of strategies to "teach to the test" (Cuban, 2013). These strategies included (a) teaching only the grade-level content that was assessed on the state test (McMurrer, 2007; Perlstein, 2010), (b) using released items from state tests (Reback, Rockoff, & Schwartz, 2011; Ryan, 2010; Srikantaiah, 2009), and (c) shifting time during the school day towards the tested subjects and away from non-tested subjects, such as science, social studies, and electives (Darling-Hammond, 2007; Dee, Jacob, & Schwartz, 2012; McMurrer, 2007; Murnane & Papay, 2010; Reback et al., 2011). In addition, in the early years of NCLB, the test scores of certain special education students did not count towards their school's rating. Schools responded to the incentives of this rule by increasing the number of students identified as special education after the policy was enacted (Cullen & Reback, 2006; Figlio & Getzler, 2006; Jacob, 2005). There is also evidence that schools selectively disciplined low-performing students more harshly around testing time to prevent them from being part of the testing pool and the resulting school rating (Figlio, 2006). These behaviors suggest that schools were aware of how they were held accountable and responded to the incentives created by those metrics. Campbell (1979) would have predicted this response under a high-stakes testing regime such as NCLB, as he posited that "the more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (p. 85).

Because NCLB attached sanctions to student proficiency levels and gave no credit for student growth, schools under accountability pressure may have responded by engaging in triage. In educational triage, schools classify students into three groups based on their perceived probability of meeting proficiency: (a) students who are unlikely to meet the proficiency standard that year are considered too low to help (called “lost causes” by a Texas elementary teacher [Booher-Jennings, 2005, p. 41]), (b) students who are predicted to pass without additional resources are considered “safe” cases, and (c) students who are close to passing are considered “suitable” cases who might meet proficiency with some additional support (i.e., the bubble students). After dividing students into three groups, schools then divert resources away from low- and high-performers and towards the suitable cases (Gillborn & Youdell, 1999).

In summary, educational triage involves (a) sorting and labeling students and (b) allocating resources to certain students at the expense of others. These actions are considered gaming behaviors because schools decide which students should receive additional support based on who would most improve overall school performance metrics rather than based on the individual needs of the students.

Empirical Background

Educational triage. Qualitative research provides accounts of how educational triage plays out in schools. Drawing on observations, interviews, and documents, Booher-Jennings (2005) reports that teachers in a Texas elementary school, at the directive of the principal and district, used benchmark test data to classify students into three groups and then diverted resources to the bubble students. Similarly, Horn (2016) uses video-recordings of middle school math teacher meetings to investigate how benchmark data inform teacher lesson planning. In one school, teachers identified which students they would be able to "get over the [proficiency] hump" by using the benchmark scores and test-score labels (Horn, 2016, p. 15). In both cases, teachers used benchmark data to identify the students close to proficiency and then provided interventions only for those students. The resources that were offered to the bubble students—but not provided for lower-achieving students—included extra attention during class, preferential seating, pull-out interventions during the school day, after-school and Saturday tutoring, and additional test preparation activities (Booher-Jennings, 2005; Diamond & Spillane, 2004; Horn, 2016; Horn, Kane, & Wilson, 2015). These studies show that some schools used benchmark data to sort students and then focused resources inequitably on bubble students.

These reports indicate that schools engaged in triage, and additional research suggests that the practice may occur somewhat frequently. A large majority of principals, along with nearly 25% of elementary and middle school teachers surveyed in three states, report that they focused more on students close to proficiency in response to high-stakes testing (Hamilton, Berends, & Stecher, 2005). In addition, more than half of the teachers surveyed agreed that high-performing students were not appropriately challenged (Hamilton, et al., 2005). Worth noting about these findings is that it is not only teachers who engage in these activities, but also school leaders who encourage and support these practices (e.g., Brown & Clift, 2010). For instance, a study of Philadelphia Public Schools' benchmark system indicates that district leaders expected low-performing schools to use test data to "identify, support, and monitor students who were close to proficiency" (Bulkley, Christman, Goertz, & Lawrence, 2010, p. 196). In all of these cases, these schools used data provided through district benchmark assessments to engage in educational triage.

These studies clearly demonstrate that some schools behave inequitably when they are held accountable for the percent of students who meet the proficiency threshold. Do these small-scale and case studies represent relatively isolated events, or are they indicative of more wide-spread responses to accountability pressure? Quantitative researchers have used large datasets to examine this question.

A challenge of using administrative data to investigate triage is identifying the bubble students: which students would schools consider "on the bubble" of proficiency? Previous studies generally categorize bubble students based on either their previous year's state test score (e.g., Ladd & Lauen, 2010) or a low-stakes exam that does not influence school ratings (e.g., Krieg, 2008). One strategy prior researchers employ to identify these students is to divide the test score distribution into quantiles, with students in the middle quantiles considered bubble students (e.g., Dee & Jacob, 2011; Neal & Schanzenbach, 2010). Another common strategy classifies students based on their distance from proficiency. For instance, Lauen and colleagues classify students as "bubble" if they score one-half of a standard deviation above or below the proficiency line (e.g., Ladd & Lauen, 2010; Lauen & Gaddis, 2012). Researchers do not know who the bubble students are, and in all of these studies, the researchers made arbitrary decisions which may be inaccurate.

The problem facing researchers, however, is the same problem facing schools. If a principal or teacher wanted to know which students might meet proficiency with some additional resources, how would they decide which students to select? Who are the bubble students? The qualitative case studies just described indicate that educators use results from local benchmark assessments to identify these students, information that is not typically included in administrative datasets. The local assessment information employed by this study represents a contribution to the triage literature because the benchmarks provide schools with the most up-to-date indicators of student mastery of the current grade-level standards, which is what the state test will cover. Because of this, schools may have used the benchmark labels to identify their bubble students. Furthermore, schools received this information three times per year, which allows for exploring whether the benchmark data had different effects over the course of the school year (something that is unique to this study). The quantitative analyses in Chapters 2 and 3 investigate the degree to which these labels impact student outcomes and whether the effects are consistent with educational triage.

Because administrative datasets generally lack measures of resource allocation, researchers rely on patterns of test score increases as proxies for strategic behavior (Springer, 2008b). As previously noted, researchers generally use one of two strategies to identify bubble students in the administrative data. After students are classified, researchers generally compare test score gains between these groups to see whether those in the middle gain more than those at the ends before and after high-stakes accountability systems like NCLB were implemented. The body of work investigating triage has found mixed results. Some research finds that under high-stakes accountability systems, bubble students make larger-than-expected gains while low- and high-performing students make smaller-than-expected gains (Jennings & Sohn, 2014; Krieg, 2008; Neal & Schanzenbach, 2010) which supports the educational triage theory. Jennings and Sohn (2014) and Neal and Schanzenbach (2010) both use data from large urban districts (Houston and Chicago, respectively) to estimate gains between low- and high-stakes tests for students across deciles of the test score distribution before and after NCLB was introduced. These studies estimate that the lowest-performing students perform about 0.04 to 0.11 standard deviations worse after NCLB was introduced. Krieg (2008) uses Washington state data to estimate that students close to proficiency perform about 0.02 to 0.05 standard deviations higher if their school failed AYP. Other studies find that test score increases for bubble students are accompanied by similarly-sized gains for low-performing students (Ladd & Lauen, 2010; Springer, 2008b) and no negative effect for the highest-performing students (Ballou & Springer, 2016; Reback, 2008), suggesting that NCLB had its intended effect of supporting the academic achievement of all students.

I offer a model in Figure 1 that represents the implicit logic in quantitative work using patterns of test score gains as evidence of triage by schools. The underlying rationale in this work is that schools will identify students who have the largest impact on their school rating (in this case, bubble students) and then allocate resources to those students. If these additional resources are effective, then students who receive them would have larger test score gains than students who do not receive these resources. The middle two steps in Figure 1 (where resources are allocated and those resources are effective) are in a black box to highlight the fact that quantitative analyses using test scores as proxies suffer from the “black box” problem.

An open question is why previous quantitative research has often failed to detect a focus on bubble students. The logic model in Figure 1 points to some possible reasons. First, researchers and schools may not agree on who the bubble students are. As already noted, case studies that detail triage indicate that schools use benchmarks to identify their bubble students (Booher-Jennings, 2005; Horn et al., 2015). Because most administrative datasets lack this level of information, researchers may misidentify the bubble students. If this were the case, prior researchers are finding null results because they are not looking at the students who actually received special treatment. A second reason for null results is that schools may not be targeting the bubble students. If schools were behaving more equitably by supporting students across the test score distribution, then there would be no differential outcomes for low-, high-, and middle-performers. A third explanation is that schools are engaging in triage but are not very good at it. This would mean that bubble students were targeted but that the special treatment did not improve test scores. If schools are allocating resources in ways that are not effective in raising test scores, this would be concerning from an efficiency standpoint. It would also be concerning from a research perspective because if this strategic behavior is not detectable in student outcomes, then test scores are not good proxies for resource allocation. While the quantitative analyses in this dissertation similarly suffer from this issue, the qualitative work reported in Chapter 4 removes the need for proxies by analyzing school documents. These documents show how school leaders proposed to use additional funds, including what resources were provided to which students. This information helps illuminate the black box in Figure 1 regarding how schools used the benchmark data, which has implications for policymakers, district leaders, and researchers.

Labeling students. Labeling students is not problematic if the labels are used to provide students with instruction at their needed level (i.e., students who have already mastered the grade-level standards are offered challenging content, while those have did not receive remediation in their areas of need). The labels are problematic, however, if they facilitate triage.

For instance, Horn (2016) reported on a school that used benchmark labels to divide students into triage groups. The seventh-grade teachers discussed in detail the all-day math camp that they were planning for their bubble students. One teacher spoke about "the kids that are gonna be left behind…are the really, really struggling children and then your kids that don't need camp" (Horn, 2016, p. 17). This teacher is describing students from the “too low” and “safe” triage groups, respectively. These teachers identified the students who were excluded from this intervention based on the district-supplied test-score labels. The math teachers taught bubble students at the all-day math camp (meaning these students received small group instruction from their math teachers), while the students who were left behind were taught by substitute teachers. This illustrates triage, where resources were diverted away from low- and high-performers and towards bubble students. While high-performing students would not need the grade-level remediation offered to the bubble students, the lowest-performing students certainly require additional support. The exclusion of the low-achievers from the learning opportunities provided to the bubble students highlights the inequities of triage.

Papay, Murnane, & Willett (2011) examine the effects of test-score labels on college enrollment using Massachusetts data. Massachusetts divides the state test score distribution into four regions, similar to the state used in this analysis. Using a regression discontinuity design, Papay and colleagues (2011) compare the college enrollment rates of students who fell on opposite sides of a performance category on their eighth-grade state test. The results indicate that the eighth-grade label mattered even though it provided no additional information beyond the test score. The researchers offer as one possible explanation that teachers may base "decisions and behaviors on the performance label and not the underlying test score" (Papay et al., 2011, p. 28). This hypothesis is investigated in Chapters 2 and 3. The authors found that the labeling effects were strongest for students who received the highest and lowest labels of the four groups with no difference in outcomes for students in the two groups around the proficiency line. They posit that the differences between these groups around the proficiency line would be minimized if schools focused on bubble students (Papay et al., 2011).

These studies suggest that teachers may use test-score labels as short-cut summaries of student ability when they decide which students should get additional support. Under a high-stakes accountability system where schools feel pressure to meet required proficiency rates, these benchmark labels may enable triage by identifying which students are closest to passing. An open question is the extent to which the accountability pressures changed with the adoption of the waiver accountability system and whether the incentive changes mitigated strategic behaviors that may have resulted from NCLB.

State Accountability Context

NCLB’s incentives were relatively clear for schools: get a certain percentage of students to meet proficiency (called Adequate Yearly Progress [AYP]) or face consequences. These consequences ranged from allowing students intra-district transfer to successful schools (called NCLB school choice) to restructuring the school and possibly firing all of the staff (Manna, 2011). Implemented beginning in 2003-04, NCLB required that the percentage of students meeting proficiency increase each year until 2013-14, when 100% of students had to meet proficiency. NCLB was primarily directed at schools: the school received the AYP rating and any consequences stemming from failing to meet AYP. Students who did not meet proficiency generally faced no consequences (although there were some states that did not allow students in certain grades to be promoted to the next grade until they met proficiency) (Manna, 2011). Proficiency rates did not directly impact teachers or principals until the sixth consecutive year of failing to meet AYP, when the state could take over these continuously low-performing schools, and teachers and administrators could lose their jobs.

In 2011, because Congress failed to reauthorize NCLB, President Obama offered states the opportunity to apply for waivers from some of NCLB's requirements. To receive a waiver, states submitted proposals that detailed their new accountability systems. Throughout their waiver application, this state's policymakers laid out their beliefs and intentions for the accountability system in light of lessons learned from NCLB. For instance, because more than half of the state's schools failed to meet AYP in 2011-12, NCLB's 100% proficiency requirement by 2013-14 was described as "unrealistic and de-motivating" (Elementary and Secondary Education Act [ESEA] Flexibility Request , 2012, p. 34).

The waiver, which was implemented in this state beginning in the 2012-13 school year, introduced a suite of policy changes for districts, schools, administrators, and teachers. The waiver proposal describes an accountability system with clear equity intentions and a focus on growth: "we premise our goals on growth against the current baseline…we do believe that all students, classes, schools and LEAs [local education agencies] have equal capacity to improve against their current baseline" (ESEA Flexibility Request, 2012, p. 34). Policymakers wanted to see growth against current performance across the education system. The same section of the document explained that "[t]his focus on growth against our current performance level meets each child, teacher, principal and LEA superintendent in the right place and creates accountability that is fair but ambitious" (ESEA Flexibility Request, 2012, p. 34). The state intended to create a fair accountability system with a "focus on growing every student, every year" (ESEA Flexibility Request, 2012, p. 43). With that focus, districts and schools were responsible for two primary goals: (a) growth for all students, every year, and (b) closing achievement gaps. The state had a two-pronged approach for meeting those accountability goals. The first was through the district and school accountability system, and the second was through the teacher and principal evaluation framework. Did the waiver's incentives match these intentions?

District and School Accountability System

Under the waiver, the state no longer directly held schools accountable. This is because policymakers did not believe that "direct state intervention in schools generally is an effective strategy for driving improvement" (ESEA Flexibility Request, 2012, p. 34). Instead, the state held LEAs accountable, and LEAs held schools accountable, for student achievement. The exception was in the state's lowest-performing 5% of schools, which remained under state oversight. The vast majority of schools did not face prescribed consequences for failing to meet AYP because (a) there were no longer set AYP requirements and (b) the LEA was responsible for intervening in schools that the LEA (not the state) identified as needing support.

During the waiver, the state evaluated districts and low-performing schools on two different measures—overall achievement and gap closure. Both overall achievement and gap closure were primarily measured by proficiency rates. Like NCLB, the overall achievement measure was simply based on the percentage of students scoring proficient. A difference from NCLB is that the required proficiency rate was no longer the same for districts and schools across the state. The required proficiency rates were calculated from the LEA's current baseline, and the state called for "each LEA to have targets of advancing proficiency levels at a steady and ambitious rate over the next four years, and for our LEAs to ask all schools to do the same" (ESEA Flexibility Request, 2012, p. 34). Gap closure was measured by reducing the difference in proficiency rates between students from historically underperforming subgroups (i.e., students in racial/ethnic sub-groups that perform below the state average, economically disadvantaged students, students with disabilities and English Learners). Although students from these subgroups must "grow proficiency levels faster than other students" (ESEA Flexibility Request, 2012, p. 34), this performance indicator measures growth in proficiency rates rather than year-to-year growth for individual students. Despite the policymakers' intentions to maintain a focus on "growth for all students, every year," the proficiency metrics assessing that goal continued to incentivize schools to focus on bubble students.

The waiver changed the students for which districts and schools were held accountable. Under NCLB, school ratings were calculated based on the aggregate proficiency rates for all students in the school from third to eighth grade. Under the waiver, the state rated schools based on students (a) in third grade math and reading, (b) in seventh grade math and reading, and (c) in aggregate grades three to eight in math and reading. Students in third and seventh grade were included both in the separate grade measures and in the aggregate, meaning they counted twice for school ratings. The waiver, then, created an incentive to focus on third and seventh grade students.

Educator Evaluation System

Through their waiver, the state continued to implement a state-wide student outcomes-based educator evaluation system that was first introduced in 2011-12. The waiver indicates that the "teacher and principal evaluation framework uses student growth through value-added scores, ensuring that across the state, we maintain a focus on advancing each child against the current baseline results" (ESEA Flexibility Request, 2012, p. 34). Teachers of tested subjects (i.e., math and reading), had 35% of their final effectiveness rating based on their students' value-added scores. For principals, 35% of their evaluation is based on school-wide growth data[1]. This is in contrast with the previous educator evaluation system, which relied on observations to rate principals and teachers but which did not take into account student achievement. The addition of growth metrics to the evaluation system may reduce the incentive for educators to focus on bubble students. Because the new educator evaluation system was introduced in 2011-12, prior to the adoption of the waiver, Chapter 3 explores the possibility that schools changed behavior because the incentives in the evaluation system shifted rather than the incentives of the waiver accountability system as a whole.

To summarize, NCLB rated schools primarily on proficiency rates without accounting for student or school baseline performance. NCLB's sanctions were intended to induce schools to improve. The combination of these factors may have resulted in unintended behaviors such as educational triage. The removal of the sanctions under the waiver may have lessened the pressure to triage. Although the waiver system continued to hold districts and schools accountable mainly for proficiency rates, the teacher and principal evaluation systems included a value-added growth component. These two metrics created competing incentives for educators. Because the GCPS benchmark data span these shifting accountability systems, Chapters 2 and 3 explore whether these policy changes induced schools to change their behavior towards low-performing students.

District Context: Greenfield County Public Schools

After a new state test in 2008 indicated that a large proportion of their students scored below proficiency, GCPS implemented a benchmark assessment policy to inform schools about student performance throughout the school year. For the 2008-09 school year, the district

[1] Both teachers and principals had 15% of their evaluation scores based on overall student achievement (i.e., proficiency rates).

adopted the Discovery Education Assessment (DEA) benchmark tests, which were "designed to predict student performance on the next high-stakes test the student will experience" (Discovery Education Assessment Research, n.d., p. 3). These predictions came in the form of test-score labels assigned to students based on their benchmark scores (either Advanced, Proficient, Basic, or Below Basic) which represent the category in which students were projected to score on the end-of-year state test later in the spring. Students in third through eighth grade took the DEA benchmarks three times per year in math and reading (in September, November, and February).

After each benchmark, the number of questions each student answered correctly and their corresponding performance labels were shared with schools (i.e., school leaders and teachers) through the district's digital data warehouse. In addition to sharing this information, GCPS helped schools access, understand, and utilize the data to support student learning, a strategy which included hiring data coaches to work with schools. This large investment in their benchmark policy makes GCPS similar to other urban districts. A survey found that 82% of the largest urban districts in the country invested in interim assessment technology (Burch, 2010). In addition, the coaching supports and technology provided by GCPS are similar to those described by other district leaders who implemented benchmark systems (Davidson & Frohbieter, 2011).

The GCPS research team recognized the pressure schools felt under the state accountability system and acknowledged that schools under accountability pressure focused on bubble students before the state test. Although data on student performance was shared with schools throughout the school year, district leaders were concerned that schools might focus exclusively on bubble students. They shared GCPS data with school leaders showing that students could make tremendous growth over the course of a year and encouraged schools not to target bubble students too early. GCPS supplied the benchmark scores and labels from 2009-10 through 2013-14. Chapter 2 represents a contribution to the literature regarding educational triage because it explores the extent to which the benchmark data may have been used differently by schools throughout the school year, information that has not been utilized in prior research.

In February of 2013 (the year the state implemented the waiver accountability system), GCPS provided schools even more prescriptive test-score labels. In this year, the state electronically shared projections for each student based on their previous test history. The GCPS research team combined these state projections with current benchmark performance to identify “priority” students. Priority students were identified through a document shared with school leaders entitled *Identification of Target Students*. This document, recreated in Figure 2, classified students into four groups based on their probability of passing the state test: Multi-Year Plan (MYP), Priority 1, Priority 2, and Enrichment. The Priority students were described in this document as students who might meet proficiency with additional support. The Enrichment students were described as having "very high success" on state tests, and the MYP students were described as having "very low success." The labels, descriptions, and color-coding (red for MYP and green for Enrichment) found in this document are aligned with the triage hypothesis, with Priority students representing the district-identified bubble students who are closest to meeting proficiency, the MYP students are those who are too low to reach proficiency this year, and the Enrichment students are the safe cases.

Although they cautioned schools that a focus on Priority students was a short-term solution to increase proficiency rates, district leaders encouraged schools to target these students as they prepared for state tests. This information is aligned with the case studies that show local assessments (and not prior year test scores) are used to determine who the bubble students are

(Booher-Jennings, 2005; Horn, 2016). Chapter 3 examines the effect of these more prescriptive labels on student outcomes.

Around the same time that they shared the Priority label with schools in 2012-13, the district selected 29 low-performing schools to receive additional funding for remediation before the state test. Each of the 29 school leaders completed a Targeted Academic Intervention Proposal (TAIP) which described their plans for restructuring the school day to provide targeted interventions. The district encouraged schools to target Priority students with the TAIP interventions, although schools were given leeway in spending the funds. These proposals from a large number of schools are analyzed in Chapter 4 and offer a view into the different strategies designed by school leaders to increase student test scores. Because this TAIP funding was offered to low-performing schools during the waiver, the TAIP documents provide insight into how schools behaved after the accountability incentives changed.

Figures

Figure 1. Logic model of how targeting resources to marginal students would lead to increased test scores

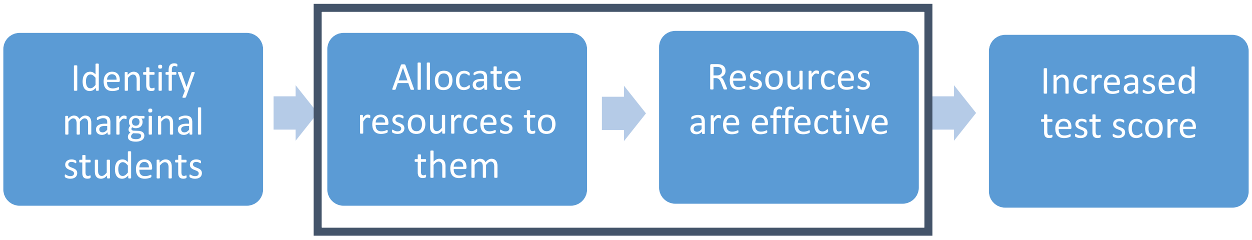


Figure 2. District training document: *Identification of Target Students*

Identification of *Target* Students

- Run the *Virtual Data Wall Report* in the *Assessment* folder of the data warehouse
- Prioritize *Target* students utilizing the summary table at the top of the report (State Projections vs most recent benchmark results) and the recommendations below:

		Benchmark			
		Advanced	**Proficient**	**Basic**	**Below Basic**
State Projection	**Advanced**	**Enrichment** *Moderate Numbers, Very High Success*	**Enrichment** *Low Numbers, Very High Success*	*Very Low Numbers, High Success*	*Very Low Numbers, High Success*
	Proficient	**Enrichment** *Moderate Numbers, Very High Success*	**Priority 2** (Benchmark in lower Proficient range*) *High Numbers, High Success*	**Priority 1** *Moderate Numbers, Moderate Success*	*Very Low Numbers, Moderate Success*
	Basic	**Priority 2** *Moderate Numbers, High Success*	**Priority 1** *High Numbers, Moderate Success*	**Priority 2** (Benchmark in upper Basic range*) *High Numbers, Low Success*	**Multi-Year Plan** *Moderate Numbers, Very Low Success*
	Below Basic	*Very Low Numbers, Low Success*	**Priority 2** *Low Numbers, Low Success*	**Multi-Year Plan** *Moderate Numbers, Very Low Success*	**Multi-Year Plan** *High Numbers, Very Low Success*
	Missing	**Enrichment** *Very High Success*	**Priority 2** (Benchmark in lower Proficient range*) *High Success*	**Priority 1** (Benchmark in upper Basic range*) *Low Success*	**Multi-Year Plan** *Very Low Success*

* Upper Basic and Lower Proficient range defined as within a few (3-4) items of number correct cut score for Proficient level.

Chapter 2: Did the Adoption of the Waiver Accountability System Mitigate Triage Behavior?

The overarching purpose of this chapter is to explore the extent to which benchmark test-score labels assigned to students affect student outcomes. This analysis examines the differential effects of the performance labels yielded by the three yearly benchmarks while taking account of incentive changes associated with the adoption of the waiver. Using a combination of regression discontinuity and difference-in-differences strategies, the labeling effects are calculated for students whose benchmark scores place them close to the thresholds separating the labels of Below Basic, Basic, Proficient, and Advanced. This chapter investigates whether schools used the district-supplied labels to triage—with the hypothesis that students close to proficiency will show more improvement than low- or high-performers during NCLB—along with the possibility that school behavior changed after the adoption of the waiver.

As described in the Introduction, much prior research on educational triage typically utilizes students' prior year test scores to determine which students schools would consider "on the bubble" of proficiency (e.g., Ladd & Lauen, 2010; Reback, 2008). There are several reasons why schools would likely employ benchmark scores rather than prior year test scores to do this. Benchmark scores represent the most current snapshot of student performance and are updated three times per year. The benchmarks assess each student's readiness for meeting the current year's grade level standards and are used to project how students will perform on the state test. Prior year test scores indicate student mastery of the previous year's tested standards but would not necessarily offer information regarding student understanding of current standards. Benchmark scores would also reflect any summer learning loss. Furthermore, case studies of educational triage indicate that schools engage in triage using benchmark results, not prior year test scores (e.g., Horn, 2016). This analysis benefits from the rich amount of district-level testing data shared with schools at different points in the school year and allows for making causal inferences about the effects of the label, something prior triage work has not been able to do.

Prior triage studies have not investigated differences between elementary and middle schools. There are reasons to believe schools might respond differently to the student testing information, and it is worth exploring whether schools vary in how they use benchmark data. For instance, the structure of the school day often has elementary teachers provide instruction for all core subjects whereas middle school teachers generally offer instruction in only one content area. Because elementary teachers spend much of the day with their students (rather than the single period that most middle school teachers have with students), it may be more difficult for elementary teachers to engage in triage and write off a group of students whom they know well as “too low” to receive support. Middle schools, on the other hand, may be better able to shift students to alternate classes to provide additional remediation for bubble students because students are used to changing classes and teachers. Additionally, middle schools generally have larger and more diverse student populations than elementary schools, which could influence the accountability pressure schools felt (Balfanz, Legters, West, & Weber, 2007).

The specific research questions explored in this chapter are:

- *To what extent do the benchmark performance labels assigned to students in September, November, and February affect end-of-year test scores during NCLB?*
- *To what degree did these relationships change after the state implemented a waiver accountability system?*
- *Do these relationships differ for elementary and middle schools?*

Data and Methods

Sample

This analysis uses student-level math and reading test scores from the 2009-10 through 2013-14 school years for students in Greenfield County Public Schools (GCPS). The benchmarks were administered to elementary and middle school students in third to eighth grade three times per year. There are about 30,000 students in these grades each year from 2009-10 to 2013-14, meaning the entire GCPS sample includes about 155,000 total observations. As described in the Methods section, each analysis limits the sample of students to those who are close to the cut-score for each test-score label.

Data

Dependent variables. The outcome of interest is each individual student's end-of-year state test score in math and reading. The test scores were standardized at the year-grade-subject level, meaning they represent the number of standard deviations above or below the mean of all other students in the state who were in the same grade and took the same subject-area test that year.

Independent variables. The three Discovery Education Assessment (DEA) benchmark assessments represent three different independent variables. In each school year, Benchmark A was administered in September, Benchmark B was administered in November, and Benchmark C was administered in February. Each benchmark score represents the number of questions answered correctly on that exam, ranging from 0 to 33 for math and from 0 to 40 for reading.

The test-score labels for these assessments comprise the main variables of interest (i.e., "treatment"). On each of these assessments, students were labeled as either Advanced, Proficient, Basic, or Below Basic based on the number of questions they answered correctly. The threshold for each performance label varied across each of the three benchmarks, two subjects, and even by grade level and year[2]. In each analysis, as detailed in the next section, each student's benchmark raw score was centered on the specific cut-score for their grade/subject/year for the threshold being analyzed to take into account this variation across tests.

Control variables. The data shared by GCPS contain background information about students. The student-level variables, employed to increase the precision of estimates, include race/ethnicity, prior year state test score, free and reduced-price lunch (FRPL) status, disability status, and whether the student was an English language learner (ELL).

School-level variables were also included to control for time-varying factors that may influence student outcomes. These include the log enrollment of the school and the percentage of students who are Black, Hispanic, White, of another race, and eligible for FRPL (gathered from the Common Core of Data for each school year). The percentage of students in each school who have a disability or are an ELL was calculated based on the GCPS data. Each school's prior year proficiency rate in math and reading came from the state's accountability website.

Accountability eras. A binary indicator *Waiver* represents which school accountability system was in place during each school year. A value of zero represents NCLB (2009-10 through 2011-12), and a value of one represents the waiver (2012-13 through 2013-14).

Methods

Because benchmark performance labels are determined by students' benchmark scores in relation to specific cut-scores on a continuous measure, regression discontinuity (RD) is a strong

[2] The cut-scores were not consistent over time because the DEA determined the cut-scores after each benchmark based on student performance.

method for assessing the causal effect of the label (Imbens & Lemieux, 2008; Shadish, Cook, & Campbell, 2002). This analysis represents a sharp RD design because all students who score above the cut-points are assigned the same performance label (Imbens & Lemieux, 2008). The RD analyses compare outcomes for students whose benchmark scores place them close to the threshold separating two different performance labels.

The hypothesis being tested is that schools used certain labels as shortcuts to determine which students should receive special treatment by engaging in triage. Under this hypothesis, students labeled Basic are likely viewed as bubble students because that is the label assigned to students who are not projected to meet proficiency, with the Below Basic label representing students who are too low to meet proficiency and the Advanced label representing students who are unlikely to drop below proficiency (i.e., the "safe" cases). For there to be a label effect, schools would need to treat students differently whose test scores were almost identical. It is worth noting that schools could have practiced triage using other information to identify the bubble students, a conjecture that is explored in the next chapter. Conversations with district personnel indicate that many schools did, in fact, use these labels to determine which students to target with additional resources. Additional information supporting the assumption that schools use these performance labels will also be forthcoming in Chapter 4.

The magnitude and statistical significance of differences in outcomes are estimated for students close to each test-score label threshold on each of the three benchmarks using local linear regression. This estimates the local average treatment effect for students with test scores close to the threshold for each performance label (Imbens & Kalyanaraman, 2009). To estimate the discontinuities in outcomes, first, the sample is limited only to students for the particular benchmark whose scores place them in the two performance labels on either side of a threshold (i.e., separately for students who were labeled as (a) Below Basic or Basic, (b) Basic or Proficient, and (c) Proficient or Advanced)[3]. Second, the number of questions each student answered correctly is centered on the cut-score that determines whether the student received the higher of the two labels (i.e., Basic, Proficient, or Advanced, respectively). The "treatment" in each model is a binary variable indicating whether the student received the higher of the two labels. Third, ordinary least squares regression is used to calculate the discontinuity at the threshold for each performance level, which allows for flexibility in including covariates and various fixed effects.

Because this analysis explores whether there is a difference in the effect of the performance labels across accountability eras, difference-in-differences (DID) within this RD framework is used to compare the difference in the discontinuity between the NCLB and waiver accountability regimes. This is done by including an interaction term between the treatment indicator and a binary variable indicating whether the state was under the waiver accountability system in a given school year. For each benchmark and threshold, I use the least squares form shown below in Equation 2.

$$Y_{sijt} = \beta_0 + \beta_1 T_{sijt} + \beta_2\,(T_{sijt}\ x\ Waiver_t) + \beta_3\ num\_correct_{sijt} + \beta_4\,(T_{sijt}\ x\ num\_correct_{sijt}) + \delta X_{it} + \gamma Y_{jt} + \eta_j + \theta_t + e_{ijt} \qquad (1)$$

In these models, Y_{sijt} represents the standardized end-of-year state test score on the subject-area test *s* for student *i* in school *j* in year *t*. T_{sijt} represents treatment (a binary variable indicating that the student received the higher of the two performance labels in subject *s*),

[3] In the rest of this paper, the Below Basic/Basic threshold will be referred to as the Basic threshold, the Basic/Proficient threshold will be referred to as the Proficient threshold, and the Advanced/Proficient threshold will be referred to as the Advanced threshold.

$Waiver_t$ is a binary variable indicating that the state was under the waiver accountability system in year *t*, $num\_correct_{sijt}$ represents the distance from the threshold (students who score below the cut-score have negative values, students who score at or above have zero or positive values), X_{it} is a vector of student controls (for student *i* in year *t*, including prior year state test score, race, FRPL, ELL, and disability status), Y_{jt} is a vector of time-varying school controls (for school *j* in year *t*, including log enrollment, prior year's percentage of students scoring proficient in subject *s,* and the percentage of students who are Black, Hispanic, White, FRPL, ELL, and have a disability), η_j are school fixed effects, θ_t are grade-by-year fixed effects, and e_{ijt} is the error term. Standard errors in this and all subsequent models are clustered at the school by grade by year level because, while the school fixed effects remove any between-school variation from the estimates, the errors for students who are in the same grade in a given school year might still be correlated. Using cluster-robust standard errors should remove any continuing correlation among students in the same grade and school (Nichols & Shaffer, 2007).

If schools used the test-score label as a shortcut to decide which students should receive resources, and if that resource allocation improved student test scores, then there should be a discontinuity in predicted outcomes at the cut-score for the performance label. The β_1 coefficient on T_{sijt} represents this discontinuity for students who received the higher label during the NCLB era. A statistically significant positive estimate for β_1 would indicate that receiving the higher performance label resulted in significantly higher test scores during NCLB. Worth noting is that β_1 is not expected to be higher for all models. For example, at the Basic threshold, the hypothesis is that the Basic students would be targeted over Below Basic students. This hypothesis would be supported by a positive estimate for β_1. At the Proficient threshold, however, the Basic students (or lower performance label) may be targeted. A negative coefficient on β_1 would provide evidence of students in the lower label being targeted. Null results would indicate that the label did not have an effect, suggesting that schools did not treat students differently based on their performance labels.

The other coefficient of interest in Equation 1 is β_2, which represents the difference in the estimated discontinuity at the threshold between NCLB and the waiver. The interpretation of β_2 depends on the β_1 estimate. If β_1 and β_2 have the same sign, then the estimated discontinuity from NCLB is enhanced after the waiver was implemented. If β_1 and β_2 have different signs, then the waiver ameliorates the earlier effect of NCLB. One main concern regarding educational triage is that the lowest-performing students are harmed because of the focus on bubble students. Because of this concern, the coefficients β_1 and β_2 at the Basic threshold are of special interest.

The β_3 and β_4 parameters represent the slope of the regression line for students who score below and above the cut-score for the performance label, respectively. These values are used to estimate the slope of the predicted performance on either side of the threshold. These estimates are expected to be positive, which would indicate that students who answer more questions correctly on the benchmark have higher end-of-year state test scores. The slope of the line is estimated separately on either side of the threshold because requiring slopes to be constant when they are actually different would affect the estimate of the discontinuity (Imbens & Lemieux, 2008).

In order to balance precision and bias in estimating the coefficients, local linear RD requires deciding how close to the threshold students need to score in order to be included in the analysis. If the discontinuity is estimated only for students very close to the threshold, the estimates will be less precise because there are fewer observations being compared. Alternatively, widening the bandwidth increases the precision but may introduce bias because the groups of students on either side of the cut-score may have more differences than simply the label assigned to them.

Determining the optimal bandwidth in these analyses is made somewhat complicated by the forcing variable (i.e., the number of questions answered correctly on the benchmark). The benchmarks do not include a large number of questions (33 in math, 40 in reading). Differences of a few items in the number answered correctly are indicative of relatively large differences in underlying performance. For instance, students who are four or five questions away from these cut-points have fairly different levels of performance. In math, students who answer four questions below the Proficient threshold on Benchmark C have an average outcome of -0.34, and 19.3% of these students meet proficiency on the state test. Students who answer four questions above that threshold have an average outcome of 0.50, and 75.4% meet proficiency[4]. Furthermore, the width of each performance label varies across years, subjects, grades, and thresholds. There are some examples where there are relatively few questions between thresholds, which limits how wide of a bandwidth could be used[5].

I used the *rd* command (Nichols, 2011) in the Stata statistical analysis software to calculate the optimal bandwidth suggested by Imbens and Kalyanaraman (2012) for each threshold and benchmark. The suggested ideal bandwidth in both math and reading ranged from 1.39 to 1.96. Due to the discrete nature of the forcing variable, this means that the ideal bandwidth includes students who answered two questions above to two questions below the

[4] Appendix Figures A1 and A2 show the relationship between the number of questions answered correctly on the benchmark and both (a) the percent of students meeting proficiency on the state test and (b) the mean standardized score on the state test for math and reading, respectively. Both sets of figures indicate that answering more benchmark questions correctly is positively related to these outcomes.

[5] For example, the Basic category ranges from 10 to 13 questions answered correctly for 5th grade in 2010 (meaning a maximum bandwidth of 4 questions could be used) but from 8 to 15 questions for 3rd graders in 2011 (with a bandwidth of 8).

threshold. To investigate the stability of estimates across varying bandwidths, Equation 1 is run on the subset of observations in increasingly wide bandwidths above and below each threshold (from 2 to 6 questions). These analyses across varying bandwidths serve as sensitivity tests for the estimated coefficients at the optimal bandwidth of two[6].

To answer the third research question, the sample is separated into elementary and middle school students. The same methods just described are then employed separately on those samples.

Assumptions Required for Regression Discontinuity

A credible treatment effect in RD rests on the assumption that there should not be a discontinuous change in the outcome based on a continuous change in the forcing variable unless there was some treatment at a certain cut-score on the forcing variable (Imbens & Lemieux, 2008). Various potential threats to RD analyses can arise. For example, if there is something that makes those who score just above and just below the threshold for the Basic performance level observably different, then any subsequent differences in average outcomes may be due to this selection into treatment. A few tests and figures are suggested by methodological research (e.g., Imbens & Lemieux, 2008; McCrary, 2008) to ensure underlying assumptions are true.

Smoothness of covariates at the threshold. I checked for underlying differences between the groups on observable covariates that could drive subsequent differences in student performance, which is recommended by Imbens and Lemieux (2008). As mentioned previously, the cut-scores for each threshold on each of the three benchmarks varied across grade, year, and

[6] The narrowest bandwidth of one does not allow for estimating the slope of the regression line. Because the forcing variable is discrete, the estimated discontinuity for a bandwidth of one represents the difference in mean outcome for students who score at the threshold compared to those who missed that label by one question as opposed to the difference at the threshold (i.e., the difference between the intercepts of the two regression lines).

subject. This variation occurred because the performance labels were assigned after the students took the assessment by the company administering the DEA. Because the thresholds were externally determined after students took the test, it is highly unlikely that there would be significant differences by prior test score, race, FRPL, ELL, or disability status around the cut-score. The smoothness of covariates at the threshold was checked by using each student covariate as the dependent variable in a simplified form of Equation 1 shown here:

$$Y_{sijt} = \beta_0 + \beta_1 T_{sijt} + \beta_2\ num\_correct_{sijt} + \beta_3(T_{sijt}\ x\ num\_correct_{sijt}) + \delta X_{it} + \gamma Y_{jt} + \eta_j + \theta_t + e_{ijt} \quad (2)$$

The discontinuity (β_1) is estimated using Equation 2 at each threshold across varying bandwidths around the threshold separately for the NCLB and waiver eras. Results for these 252 tests (two subjects x two eras x three thresholds x three benchmarks x seven student covariates, not shown) indicate no consistent significant difference in covariates around any of the thresholds in either era.

Manipulation of the running variable. Identification of a discontinuity in the regression function requires an assumption of continuity in the density of the number of questions answered correctly around each threshold (McCrary, 2008). If agents are able to manipulate values on the running variable (that is, manipulate the number of correctly-answered questions and the subsequent performance label), then resulting differences in outcomes may be based on this selection. For example, this could occur if teachers systematically changed some students' answer sheets so instead of being labeled Below Basic, students would receive the Basic label (i.e., treatment of receiving the higher of the two labels). If the students whose scores were adjusted would have scored higher on the state test in absence of this manipulation, these actions could create an artificial discontinuity at that threshold. Manipulation of benchmark test-score labels is less plausible in this case because the performance label cut-scores are externally calculated after students take the test, a situation which meets the criteria set by the What Works Clearinghouse regarding the integrity of the forcing variable in RD analyses (Schochet et al., 2010). I tested for manipulation visually through a series of histograms plotting the number of questions answered correctly around each threshold for each benchmark (Imbens & Lemiux, 2008; McCrary, 2008). Neither subject shows evidence of manipulation around any thresholds (shown in Appendix Figures A3 for math and A4 for reading).

Results

Descriptive Statistics

The percentage of students classified as Below Basic, Basic, Proficient, and Advanced on each benchmark during the two accountability eras are shown in Figure 3 (math) and Figure 4 (reading). These graphics indicate that the labels were more consistent across the benchmarks during NCLB than the waiver. For example, about 23% of students were labeled Below Basic on each math benchmark during NCLB, whereas that percentage dropped ten points between Benchmark A (23%) and Benchmark C (13%) during the waiver. Nearly 40% of the students are labeled Basic and about 28% are labeled Proficient in math during both eras. Figure 4 shows that all three reading benchmarks had similar breakdowns by label: about 15% of students were labeled Below Basic, 39% as Basic, 34% as Proficient, and 12% as Advanced. During the waiver, the percentage of students labeled Below Basic decreased between the first and third reading benchmarks with similarly-sized increases in both the Basic and Proficient labels.

Tables 1 and 2 show the descriptive characteristics for students assigned each test-score label for math and reading, respectively. The first column in each table represents the full sample, and each of the three benchmarks has been broken down into the four performance labels in the next sets of columns. Overall, the average student in GCPS scores about two-tenths of a standard deviation below the state average on the end-of-year state test. The sample is 45% Black, 18% Hispanic, and 26% White. Seventy-one percent of the sample is eligible for FRPL, 12% are ELLs, and 5% are identified as students with disabilities.

Black, Hispanic, FRPL, and ELL students (a) are over-represented in the Below Basic label for both subjects, (b) have smaller percentages on the higher labels, and (c) are under-represented in the Advanced level on all tests. For example, in math, Black students—who make up 45% of GCPS students—comprise 56% of the students labeled Below Basic but only 22% of the Advanced students. Although they make up 71% of the entire sample, FRPL-eligible students comprise 88% of Below Basic students but only 35% of the Advanced students.

The full results from Equation 1 are shown in Tables 3 and 4, for math and reading, respectively. These results are estimated by including students who score two questions above to two questions below the cut-score for each threshold (i.e., the optimal bandwidth). The full results are reported to show that the estimated values for β_3 and β_4 (the slopes for the regression line to the left and to the right of each threshold) are positive, as expected. These values mean that every additional question answered correctly on the benchmark is associated with an increase of 0.02 to 0.09 standard deviations on the state test. These positive slope estimates occur across all models of the data. Because of the large number of comparisons being made in these analyses (three thresholds for each of three benchmarks for varying bandwidths of two to six questions around the threshold for two subjects), only the coefficients for *T* and *T x Waiver* are included in subsequent tables and figures.

Math

The discontinuities estimated from Equation 1 across varying bandwidths are shown in Table 5. The coefficients marked "*T*" represent β_1, the estimated discontinuity in outcome for students who received the higher of the two labels during NCLB. These results will be discussed first. The coefficients marked "*T x waiver*" represent β_2, the difference in the estimated discontinuity at that threshold during the waiver, and are discussed second.

Each of the estimated discontinuities from Table 5 are also shown graphically with 95% confidence intervals in Figure 5. The set-up of these coefficient plots matches the structure of the table: the thresholds comprise the columns, the three benchmarks make up the rows, the estimated discontinuities during NCLB appear in the top part of each plot, and the difference in estimated discontinuities during the waiver are shown in the bottom part of each plot. The visual representations in Figure 5 are helpful for assessing whether (a) the point estimates across bandwidths are similar to one another and (b) the confidence intervals cross the zero line, indicating that the coefficient is not statistically significant. In the results, I look for consistent evidence of differences in average outcomes between groups of students on either side of the threshold. By consistent, I mean that the point estimates across the varying bandwidths are similar in magnitude and statistically significant (or very close).

No Child Left Behind (2009-10 through 2011-12). The NCLB results are shown in the top part of each panel in Table 5 and Figure 5 (*T*). The most consistent results during this era are shown in the bottom left set of results, representing the effect of the Basic label after the third math benchmark (administered in February of the school year). During NCLB, students who were barely labeled Basic scored significantly higher than students who were barely labeled Below Basic (with discontinuities ranging from 0.033 to 0.047, $p<0.01$). The larger gains for Basic students suggest that schools focused more attention on these students at the expense of Below Basic students during NCLB. This is aligned with the triage theory that schools would shift resources to students closer to meeting proficiency.

There are few other consistent or significant estimated discontinuities found on other math benchmarks or thresholds during NCLB. The discontinuities estimated at the Proficient threshold on Benchmark C, for instance, are consistently negative across bandwidths. Taken at face value, this indicates that students who are barely labeled Proficient gain less than students who are barely labeled Basic. The estimates are small in magnitude and insignificant, however, providing only limited support to the idea that schools focused more on students labeled Basic than on those labeled Proficient. At the Advanced threshold for Benchmark B, the estimated discontinuities are consistently positive but larger in magnitude and significant only at the narrowest bandwidth (β= 0.054, $p<0.05$). This is suggestive that students who barely received the Advanced label in November scored higher than students who barely received the Proficient label. There are significant and positive discontinuities at the Basic threshold on Benchmark B and at the Proficient threshold on Benchmarks A and B at the wider bandwidths, but these estimates are negative at the optimal bandwidth of two. Because these estimates are not consistent across bandwidths and are significant only when comparing students who answered 10 or 12 questions differently on a 33-question test, there are few conclusions to be drawn about school behavior.

In summary, there is some evidence that schools viewed students labeled Basic on the third math benchmark as bubble students and targeted resources to them during NCLB. Students labeled Basic who were close to either the Basic or Proficient thresholds gained more than students who were close to but on the other sides of those thresholds (although the gains are statistically significant only at the Basic threshold). These gains for Basic students do not appear to come at the expense of the highest-performing students, given that Advanced students gained more than students who barely miss that label. That the labels from earlier benchmarks did not have test score effects suggests that schools did not target Basic students until later in the school year. The third benchmark was the last data that schools received from the district prior to the state test, and schools may have viewed students labeled Below Basic at that point in the school year as unlikely to meet proficiency.

Waiver (2012-13 to 2013-14). The difference in the estimated discontinuities during the waiver period are shown in the bottom part of each panel in Table 5 and Figure 5 (marked "*T x waiver*"). There are a number of consistent and significant differences between NCLB and the waiver. At the Basic threshold, the interaction of the label with the waiver is statistically significant and negative across all bandwidths on both Benchmarks B and C. Because the estimated discontinuities at this threshold were positive during NCLB, these negative differential effects indicate that students barely labeled Below Basic benefitted after the waiver was introduced. The magnitude of the differential effect is even larger on Benchmark C (β= -0.040 to -0.066, *p<0.01*) than on Benchmark B (β≈ -0.03, *p<0.05*). Whereas the significant positive estimates for β_1 indicate that students who barely received the higher Basic label had better outcomes than students who barely received the Below Basic label during NCLB, the combined coefficients under the waiver are no different from zero. This means that the Basic label did not have an effect under the waiver, suggesting that schools no longer prioritized bubble students.

This shift in focus during the waiver period to students with the lower of the two labels is also shown at the Proficient threshold. The point estimates for *T x Waiver* are consistently negative across all three benchmarks at this threshold (ranging from -0.010 to -0.035). This suggests that schools shifted resources during the waiver era towards the lower-performing (i.e., Basic) students over those labeled Proficient. These estimates, however, are generally smaller in magnitude at lower bandwidths and significant only at wider bandwidths. This increased focus during the waiver era on lower-performing students is aligned with the results found at the Basic threshold.

To summarize, schools used the math labels differently under the two accountability regimes. During NCLB, the students who were barely labeled Below Basic on the third benchmark scored significantly lower than students who were barely labeled Basic. During the waiver, schools appear to have shifted their attention to lower-achieving students. Students who score on the lower side of the Basic and Proficient thresholds on Benchmarks B and C had better outcomes during the waiver compared to similar students during NCLB. Furthermore, schools may have shifted their attention to lower-performing students during the waiver as early as November, when the second benchmark was taken. This provides evidence that the incentive shift in the accountability system changed schools' behavior, which benefitted lower-performing students mainly in math.

Elementary and middle schools. The math results estimated separately by school level are presented in Figure 6 for elementary schools and in Figure 7 for middle schools (coefficients are included in Appendix Tables A1 and A2, respectively). The results just reported indicate that the only significant positive discontinuities estimated from NCLB were at the Basic threshold on the third benchmark. These results appear to be driven by middle schools, with Figure 7 showing consistently positive significant estimates of approximately 0.05 standard deviations for β_1. The corresponding coefficient estimates in elementary schools are small in magnitude and not measurably different from zero. This suggests that during NCLB, middle schools targeted students labeled Basic in the lead up to the state test at the expense of students labeled Below Basic, which had a detectable effect on end-of-year outcomes. These differences by school level are discussed in more detail after all of the results are presented.

One difference from the district-wide results is found at the Proficient threshold in elementary schools during NCLB. The estimates of β_1 on Benchmark C are consistently negative and statistically significant at the Proficient threshold in elementary schools ($\beta \approx -0.04$, $p<0.05$), results which did not occur in middle schools. This indicates that elementary students who were barely labeled Basic scored significantly better than students barely labeled Proficient during NCLB. This suggests that elementary schools focused their attention on students labeled Basic at the expense of those labeled Proficient under NCLB.

Both elementary and middle schools exhibit the increased focus on lower-performing students at the Basic threshold during the waiver. Both school types display consistent negative coefficients for *T x waiver* at this threshold on Benchmarks B and C. The coefficients are similar in magnitude across school types and somewhat larger in magnitude on the third benchmark (estimates ranging from -0.038 to -0.071) than the second benchmark (estimates ranging from -0.026 to -0.048). The estimates are less precise in elementary schools—likely due to the smaller sample of students—which makes the estimates significant more often in the middle schools.

Middle schools appear to be driving the full results that found significant negative coefficients at the Proficient level on both Benchmark B and C during the waiver. While both school types have consistent negative estimates at this threshold, the differential effect of the Proficient label on these two benchmarks during the waiver is larger in magnitude for middle schools (β= -0.013 to -0.048) and statistically significant across most bandwidths. The results at the Basic and Proficient thresholds suggest that middle schools supported lower performing students relatively early in the year during the waiver compared to NCLB.

Reading

No Child Left Behind (2009-10 through 2011-12). The reading results are shown in Table 6, with the corresponding coefficients presented graphically in Figure 8. As with math, I first discuss the results from NCLB, which are shown in the top part of each set of estimates, before moving onto the results during the waiver.

The most consistent results in reading during NCLB are similar to those found in math: students who are barely labeled Basic on Benchmark C score higher across all models on the end-of-year state reading test than students barely labeled Below Basic. The discontinuity estimates are all positive and similar in magnitude across the varying bandwidths (ranging from 0.019 to 0.030) but are significant only at the widest bandwidths of five and six. As with the math results, this suggests that, during NCLB, schools focused more attention in the months before the state test on students labeled Basic over those labeled Below Basic. These findings align with the triage hypothesis that schools would neglect low-performing students to focus on those close to proficiency, although the evidence of this is weaker in reading than in math.

There are few other consistent or significant differences in end-of-year state reading scores based on the test-score label at other thresholds or benchmarks. The Proficient threshold on Benchmark B reveals consistent negative estimates, which suggests that students who barely scored Proficient have lower estimated outcomes than students who barely scored Basic. These discontinuities, however, are small in magnitude ($\beta \approx -0.015$) and significant only at the widest bandwidth. As with math, this is somewhat suggestive that schools focused more on students just below the proficiency line during NCLB. This evidence is relatively narrow in reading, however, because the discontinuities in outcome around the proficiency threshold are present only on the second reading benchmark. By Benchmark C, the estimated coefficients are still negative but smaller in magnitude and no longer significant.

These results provide suggestive evidence that schools in GCPS used the district-provided labels during NCLB to decide which students to target. Students labeled as Basic appear to be viewed by schools as the bubble students, and barely receiving the Below Basic label has a negative effect on students close to the Basic threshold. The effects of the labels tend to show up later in the year (after the third benchmark, administered in February), which suggests that while schools may have diverted attention from low-performers to bubble students, they did not do so for the entire school year.

Waiver (2012-13 to 2013-14). The difference in the estimated discontinuities during the waiver are shown in the bottom set of results in Table 6 and Figure 8. Once again, the main labeling effect is found at the Basic threshold on Benchmark C. The differential effect of the Basic label for reading during the waiver era is consistently negative and significant across bandwidths (β= -0.041 to -0.053, $p<0.05$). As with math, this indicates that schools shifted their attention during the waiver to the lower-performing students.

Similar to the reading results during NCLB, there are relatively few consistent or significant estimates at other thresholds or benchmarks. The point estimates for the *Basic x Waiver* indicator are consistently negative for Benchmark A, which implies that schools began shifting attention to lower-performing students earlier in the year. These results are imprecisely measured, however, and significant only at the widest bandwidth, which limits these interpretations. Because performance labels on the early reading benchmarks did not influence student outcomes, this again suggests that schools did not use these labels to target bubble students early in the school year.

To summarize, student outcomes in reading appear to be less influenced by benchmark labels than in math. There is some evidence that students who barely received the Below Basic label on the third benchmark during NCLB scored significantly lower on the end-of-year reading state test and relatively strong evidence that those students benefitted under the waiver. Taken together with the math results, this provides evidence that schools used the performance labels from the third benchmark to target Basic students (at the expense of the Below Basic students) during NCLB. The waiver ameliorated the negative effect estimated from the NCLB era for students who just missed the Basic label on the third reading benchmark.

Elementary and Middle Schools. Reading results for elementary schools are shown in Figure 9 and for middle schools in Figure 10 (with coefficients included in Appendix Tables A3 and A4, respectively). In contrast to the math findings, the reading benchmarks displayed relatively few significant results district-wide. Separating the sample by school level, however, reveals additional differences between elementary and middle schools.

The district-wide reading results were somewhat suggestive that students who were just barely labeled Basic on the third benchmark scored better than students who were just barely labeled Below Basic during NCLB. Disaggregating these results by school level indicates that middle schools were responsible for these results. Figure 10 presents consistently positive estimates at the Basic threshold on Benchmark C during NCLB in middle schools. The estimated discontinuities are similar in magnitude across bandwidths (ranging from 0.020 to 0.035) but are statistically significant only at the higher bandwidths. Although not as strong as the math results, these findings indicate that middle schools in GCPS were more likely than elementary schools to focus on Basic students over the Below Basic ones, at least for students close to the Basic threshold.

The district-wide results indicated that Below Basic students benefitted from the waiver. The differential effect of the Basic label (compared to the Below Basic label) was consistently negative and statistically significant across varying bandwidths on the third reading benchmark. Both elementary and middle schools show this consistent negative pattern at the Basic threshold. The magnitude of the estimates is larger for middle schools (β= -0.042 to -0.060) and consistently significant. The elementary estimates (β= -0.022 to -0.040) are imprecisely measured and not statistically different from zero (likely due to the smaller sample). Both elementary and middle schools, then, appear to have shifted attention to lower-performing reading students during the waiver.

There are differences in the reading results between the school levels which did not show up in the district-wide results at the Advanced threshold. First, elementary schools display consistently negative and significant estimates at this threshold on Benchmark B during the waiver (β= -0.048 to -0.071, $p<0.05$). This indicates that elementary students who were barely labeled Proficient (compared to Advanced) gained significantly more under the waiver than similar students during NCLB. This negative differential effect in elementary schools occurred only on the second reading benchmark. The third reading benchmark showed smaller negative estimates at the Advanced threshold under the waiver, with none of them measurably different from zero. Second, there is a positive effect of the Advanced label for middle school students during the waiver. Coefficients at the Advanced threshold on reading Benchmark C are consistently positive and statistically significant (β= 0.033 to 0.055, $p<0.05$). In this case, middle school students who were barely assigned the Advanced labeled gained significantly more than students who were barely assigned the Proficient label during the waiver. This suggests that middle schools' shift in attention to lower-performing students (found on Benchmark C at the

Basic and Proficient thresholds) did not harm high-performing students. It appears that the middle school students labeled as Advanced may have benefitted from that change in focus.

Both elementary and middle schools appear to have shifted their attention during the waiver to students who were lower-performing. Many of the differences in the estimated discontinuities during the waiver are negative, suggesting that lower-performing students benefitted from the suite of changes made by the waiver. Evidence for the increased attention on lower-performing students is not as strong in reading as it is in math.

Discussion

The purpose of this analysis was to explore whether the district-supplied benchmark test-score labels that were quasi-randomly assigned to students influenced their end-of-year state test scores. The previous section reported on a large number of comparisons across label thresholds, benchmarks, subjects, accountability eras, and school types. I now step back to look for patterns across these results.

No Child Left Behind Era (2009-10 through 2011-12)

This analysis investigated the extent to which schools used the performance labels as shortcuts to perform educational triage under the incentive pressures of NCLB and if so, at which point in the year the performance labels influenced outcomes.

Math. The findings indicate that schools in GCPS focused on students who were labeled Basic on the third benchmark. This was predicted under the triage hypothesis with the idea that schools would view Basic students as “on the bubble” of meeting proficiency. Both elementary and middle schools focused on Basic students in math, although they did so at different thresholds. Middle schools focused on Basic students whose scores placed them close to the

Basic threshold during NCLB. Answering just enough questions on the third benchmark to be labeled Basic (compared to Below Basic) resulted in significantly higher end-of-year test scores for middle school students. These gains suggest that middle schools provided additional resources for Basic students but not for the Below Basic students in the last months before the state test during NCLB. On the other hand, elementary schools appear to have paid more attention to Basic students close to the Proficient threshold. Elementary schools did not differentially treat students at the lower end of the Basic label (i.e., no discontinuity at the Basic threshold).

Reading. The results from reading provide limited evidence that schools focused on students labeled Basic during NCLB. Middle school students who were barely labeled Below Basic on the third reading benchmark gained less than students barely labeled Basic. As with math, this implies that middle schools used the Below Basic label—especially on the last benchmark before the state test—as a designation that students would not improve enough to meet proficiency and instead focused on Basic students. There were no differences in outcomes for elementary students who scored close to the threshold.

Takeaways from NCLB. The results demonstrate that schools did focus on bubble students during NCLB. This focus occurred more strongly in math than in reading. Although the labels provided no additional information beyond the underlying test scores, the statistically significant results indicate that schools used the district-supplied labels to make resource allocation decisions. That the performance labels on the first two benchmarks do not relate to final test scores implies that schools did not write off low-performing students earlier in the year during NCLB.

In addition, middle schools appear to be more consistent with their behavior during NCLB. The findings indicate that middle schools focused on bubble (Basic) students over lower-performing (Below Basic) students in both math and reading. This implies that middle schools were more responsive to NCLB’s accountability incentives. This might be because middle schools had larger populations with more subgroups that had to meet NCLB proficiency rates than elementary schools. This pressure may have led middle schools to engage in strategic behavior to increase the proficiency rate. The consistency in this behavior across subjects, in schools where these subjects are taught by different teachers, may also reflect middle school leaders who encourage or facilitate this behavior.

Elementary schools, on the other hand, exhibit a strategic focus on bubble students but only in math and only for Basic students close to the Proficiency threshold. Elementary schools, then, do not appear to treat lower-performing students differently based on the label. This difference by school level may be due to the fact that elementary teachers spend more time with their students. They might have a harder time considering certain students as “too low” than middle school teachers who teach a much larger number of students throughout the school day. In addition, because elementary school teachers instruct multiple content areas, they might know a student does well in reading but not in math, which might limit their belief that a certain student is too low for support. Middle school teachers, who typically teach only one content area, would not have as broad a knowledge of their students, which could make it somewhat easier to consider students too low.

While there is evidence that schools in GCPS targeted bubble students during NCLB, a triage response also involves diverting resources away from the highest-performing students. There was no measurable difference between students who scored just above or below the

Advanced threshold during NCLB. Students whose benchmark scores place them near the Advanced threshold, however, are unlikely to drop in the distribution enough to fall below proficiency. Schools may then have viewed students close to but on either side of this threshold as “safe” and treated them the same, which could help explain why students did not have different outcomes[7].

Yet this argument that schools should not treat students differently based on their labels is true for all thresholds. On its surface, it seems counterintuitive that schools would treat students differently based on these labels, given that the district shared the test scores that produced the labels. This seems especially true for students close to but above the Proficient cut-off, although the results indicate that elementary students just above and below that threshold on the third math benchmark had different outcomes. After hearing preliminary results from this analysis, district leaders gave credence to the idea that schools did, in fact, use these labels in the way hypothesized for this study. A member of the district research team, who works directly with administrators to use benchmark data in making instructional decisions, reported that educators viewed the labels as meaningful. This individual shared that some administrators in GCPS believed that students who were labeled Proficient on a benchmark meant students would meet proficiency on the state test, believing that they “don’t have to focus on those kids anymore” (meeting with GCPS research team). Some of these school leaders expressed displeasure towards the GCPS research office at the end of the year when those students did not meet proficiency on the state test. This viewpoint represents a misunderstanding of the benchmark data, and the district research team discussed how to communicate appropriate uses (and misuses) of the

[7] A different definition of bubble students is examined in the next chapter, which groups together these higher-performing students close to the Advanced threshold as the “safe” group.

student data shared with schools. This also highlights a tension that GCPS faces—along with other districts implementing a benchmark test policy—that the information they provide, coupled with the pressures placed on schools to meet proficiency requirements, may actually assist schools to engage in strategic behavior that is inequitable.

Waiver Accountability Era (2012-13 to 2013-14)

Math. During the waiver, schools in GCPS shifted focus to lower-performing students in math and did so relatively early in the year. Both elementary and middle schools demonstrate consistent negative differential effects of the test-score label at both the Basic and Proficient thresholds. The magnitude of the coefficient estimates were similar across school types. The larger standard errors in elementary schools, however, made the coefficients significant more frequently in middle schools.

This shift in focus to lower-performing students occurred as early as Benchmark B, which was administered in November of each school year. Schools might have used the labels from that benchmark to make changes in classes or teachers for the second semester, such as reassigning lower-achieving students to the best math teachers. Alternatively, schools may have used the results from the second benchmark to determine which students should receive additional resources. These issues are explored further in Chapter 4.

Reading. The waiver was also associated with a shift to lower-performing students in reading. These results, however, were concentrated only at the Basic threshold and only on the third benchmark. Both elementary and middle schools show consistent negative differential effects at the Basic threshold. This indicates that schools focused more on students labeled Below Basic on the third benchmark during the waiver than they did during NCLB. Middle schools display effects that are larger in magnitude and more frequently significant than elementary schools. This may reflect a shift in behavior from NCLB, when only middle schools exhibited an increased focus on students receiving the Basic label over those labeled Below Basic on the third reading benchmark.

Some effects emerged at the Advanced threshold when the data were analyzed separately by school level. There is a strong positive effect for middle school students who barely received the Advanced label on the third benchmark. This might reflect that students labeled Advanced on the last benchmark before the state test receive some sort of special treatment that would benefit them (e.g., enrichment activities) while the school is providing support to other lower-performing students. On the other hand, there is a strong negative effect during the waiver for elementary students who barely received the Advanced label on the second benchmark. This does not necessarily mean that Advanced students were harmed but that the students who just missed this label gained more during the waiver. This could occur, for example, if elementary schools gave additional resources for students labeled Proficient to ensure they met proficiency on the state test.

An alternative explanation for these findings is not that schools treated students differently based on the label but that the label affected students' views of themselves[8]. In their study of how eighth-grade state test labels influence college-going rates, Papay and colleagues (2011) hypothesize that students adjust their self-identity and expectations based in part on external factors such as these performance labels. For example, being labeled as Advanced might

[8] This would require that students knew their benchmark performance. When asked if schools would have shared this information with students, a member of the GCPS research team answered that it would “have varied by school and/or classroom. A school leader would likely have an expectation around sharing data with students but that doesn’t mean every single teacher would do it… Further, if data were being shared with students, then I think the label most definitely would have been shared. It’s more likely that the label was shared than the number of questions answered correctly” (email correspondence).

enhance middle school students’ views of themselves, which might increase their motivation and subsequent performance. The Advanced label had the opposite effect for elementary students. The negative results for elementary students just below the Advanced threshold could indicate that these students were motivated to work harder so they would be labeled Advanced on the state test. This might also help explain the negative effects found during NCLB for Below Basic students. It is plausible that barely scoring Below Basic on the last benchmark before the state test may depress a low-performing students' self-view, leading to lower motivation and lower test scores. This hypothesis does not seem to bear out, however, given the results of this study. If this were the reason behind the lower performance for Below Basic students, then it is hard to explain why this label would no longer have a similar effect during the waiver.

Takeaways from the Waiver. The adoption of the waiver was associated with a shift in focus to lower-performing students in both math and reading. Students barely labeled Below Basic benefit in both subjects and in both elementary and middle schools. This shift occurred at more thresholds (both Basic and Proficient) and earlier in the year (beginning after Benchmark B) in math than in reading.

Middle schools demonstrated a more consistent shift during the waiver across both subjects. Middle schools appeared more likely to focus on students above the Basic threshold during NCLB, and students in those schools had a larger change in average outcomes under the waiver than elementary students. Middle schools in GCPS, then, appear more responsive to accountability incentives. As already noted, middle schools’ larger populations may have driven them to engage in strategic behavior during NCLB. Perhaps the release from that policy’s requirements under the waiver allowed middle schools to behave more equitably. These differences by school level point to additional research which investigates factors that may contribute to these behaviors.

Implications

The implications of this analysis for GCPS district leaders were apparent after they were presented with these results. The research team discussed what to do about their current benchmark system. The state test had changed in the previous year, meaning there was no way to project student end-of-year state test performance and assign test-score labels (as had been done with the DEA assessments used in these analyses). In this first year of the new state test, the district continued to administer the benchmark tests but shared with schools only student raw scores. Due to the findings of this study, the team deliberated about whether they should assign performance labels to benchmark scores, given that schools seemed to make instructional decisions based them. While this work is representative only of GCPS, knowing that the test-score labels have unintended effects on student outcomes could be useful to consider for other large urban districts who administer interim assessments and share that data with schools.

The results of these analyses indicate that the lowest-performing students—those who were labeled Below Basic—benefitted from the waiver policy changes. The significantly higher performance for these students under the waiver wiped out the negative impact that the label created during NCLB; during the waiver, there was no significant difference in outcome between students barely labeled Below Basic and Basic. This suggests that the suite of changes during the waiver did change the incentives for schools and led to more equitable treatment of students.

These results suggest that schools changed how they responded to student data they received over the course of the school year. This has implications for researchers who are investigating educational triage and have access only to prior year test scores. These results should lead to some caution in the conclusions made when there are no differences found in outcomes. For instance, labels on Benchmark A have no measureable effect on students' average test scores, but those on Benchmark C do. If only the first data point for students was utilized in this research, the conclusion might have been that performance labels have no effect on student outcomes. Another lesson for researchers is that elementary and middle schools responded differently to the labels. Disaggregating results by school level might reveal variation in behavior that might otherwise be concealed.

If basing school ratings on proficiency rates caused schools to consider some students too low to receive needed support, this should be concerning for policymakers who intend for school accountability systems to equitably support the learning of all students. This analysis, however, is unable to disentangle the effects of policies which all changed at the same time. It is not clear whether the improved performance of the low-performing students was due to (a) the removal of required consequences for schools that failed to meet state-determined proficiency rates, or (b) oversight by the district rather than the state, or (c) the inclusion of growth metrics in the educator evaluation system. The combination of these changes appear to have benefitted these students. The incentives that changed school behavior are certainly of policy interest and worth further research with datasets that can explore these different policy changes.

Furthermore, while the waiver, which was implemented in 2012-13, overhauled the school accountability system, the state made changes in 2011-12 to the educator evaluation system that may have shifted incentives in the direction found in these results. The 2011-12 school year was during NCLB but was the first year that the state utilized value-add growth metrics as a substantial portion of teachers' and principals' evaluation scores. That means the incentives in 2011-12 were somewhat mixed. The school rating under NCLB was comprised primarily of proficiency rates and related to consequential sanctions, which would incentivize a focus on bubble students, but educators' personal evaluations were determined in part by student growth. The more equitable behavior towards lower-performing students during the waiver demonstrated in this study may reflect the growth metrics that were implemented the previous year. The next chapter explores this as one of several alternative explanations for the reduced focus on bubble students found here.

Limitations

This analysis tested the hypothesis that schools used benchmark test-score labels to allocate resources to students but faces some limitations. The discrete nature of the running variable limits the ability to use RD models to their full effect. This was illustrated by the computer program indicating that the optimal bandwidth was a fractional value. Additionally, because the slope of the regression line is estimated using clusters of students at each individual data point and because the widths are narrow for some test-score labels, the slope estimates are limited in how variable they can be. Furthermore, these results reflect only the local average treatment effect for students close to the threshold separating the labels. While the results indicate that the Below Basic students benefitted under the waiver system, this is true for students close to the threshold but not necessarily for all low-performing students. Another limitation of this work is that test scores are used as proxies for allocation of resources by schools. It is possible that schools did in fact target students earlier in the year but that the interventions were not effective in increasing test scores. This is a limitation quantitative black box studies, and Chapter 4 addresses this issue. That chapter analyzes documents completed by school leaders to learn more about what schools do to target students (and which students are actually targeted), which helps contextualize the results from these analyses.

Benefits of using data from a single district include being able to learn more about how the district used and shared student data with schools as well as receiving access to more data than is generally available in large-scale datasets. While the district shared this benchmark data with schools for five years, the district gave schools even more prescriptive labels for students based on their benchmark scores in 2012-13 (the first year of the waiver) and provided additional funding to some schools to prioritize a group of district-identified bubble students. It is possible that the change in behavior found here is because schools defined bubble students differently. In the next chapter, I examine this as one of several alternative explanations for the reduced focus on bubble students found in this study.

Tables

Table 1. Math benchmark sample statistics, 2009-10 to 2013-14

		Benchmark A				Benchmark B				Benchmark C			
	Full Sample	BB	Basic	Prof	Adv	BB	Basic	Prof	Adv	BB	Basic	Prof	Adv
	-0.21	-0.95	-0.38	0.40	1.23	-1.07	-0.43	0.30	1.11	-1.14	-0.50	0.30	1.10
Math z-score	(0.99)	(0.73)	(0.75)	(0.72)	(0.72)	(0.70)	(0.72)	(0.71)	(0.73)	(0.68)	(0.68)	(0.66)	(0.72)
Demographic information													
Black	0.45	0.56	0.48	0.37	0.22	0.56	0.48	0.39	0.24	0.57	0.49	0.39	0.26
	(0.50)	(0.50)	(0.50)	(0.48)	(0.41)	(0.50)	(0.50)	(0.49)	(0.43)	(0.50)	(0.50)	(0.49)	(0.44)
Hispanic	0.18	0.20	0.20	0.16	0.09	0.20	0.20	0.17	0.11	0.19	0.20	0.17	0.12
	(0.38)	(0.40)	(0.40)	(0.37)	(0.29)	(0.40)	(0.40)	(0.37)	(0.31)	(0.39)	(0.40)	(0.38)	(0.33)
White	0.26	0.16	0.24	0.33	0.44	0.17	0.23	0.30	0.42	0.16	0.22	0.31	0.43
	(0.44)	(0.37)	(0.43)	(0.47)	(0.50)	(0.37)	(0.42)	(0.46)	(0.49)	(0.36)	(0.41)	(0.46)	(0.50)
FRPL	0.71	0.84	0.77	0.61	0.38	0.86	0.78	0.64	0.43	0.85	0.78	0.65	0.47
	(0.45)	(0.36)	(0.42)	(0.49)	(0.48)	(0.35)	(0.42)	(0.48)	(0.49)	(0.35)	(0.41)	(0.48)	(0.50)
English language learner	0.12	0.18	0.14	0.08	0.03	0.19	0.14	0.09	0.04	0.18	0.15	0.10	0.05
	(0.33)	(0.39)	(0.35)	(0.27)	(0.17)	(0.39)	(0.35)	(0.28)	(0.20)	(0.38)	(0.36)	(0.29)	(0.22)
Students with disabilities	0.05	0.07	0.04	0.03	0.02	0.07	0.05	0.03	0.03	0.07	0.05	0.03	0.02
	(0.21)	(0.26)	(0.20)	(0.17)	(0.15)	(0.26)	(0.21)	(0.18)	(0.16)	(0.26)	(0.22)	(0.17)	(0.15)
Observations	154480	36440	58320	34380	13570	32380	56470	38450	18090	30490	54840	41160	20080

Notes: Standard deviations in parentheses. BB means Below Basic, Prof means Proficiency, and Adv means Advanced.

Table 2. Reading benchmark sample statistics, 2009-10 to 2013-14

		Benchmark A				Benchmark B				Benchmark C			
	Full Sample	BB	Basic	Prof	Adv	BB	Basic	Prof	Adv	BB	Basic	Prof	Adv
Reading z-score	-0.21	-1.22	-0.54	0.37	1.22	-1.27	-0.53	0.38	1.20	-1.31	-0.55	0.41	1.26
	(1.00)	(0.70)	(0.67)	(0.67)	(0.70)	(0.67)	(0.65)	(0.67)	(0.71)	(0.68)	(0.66)	(0.67)	(0.71)
Demographic information													
Black	0.45	0.53	0.51	0.39	0.24	0.52	0.51	0.39	0.24	0.52	0.51	0.38	0.23
	(0.50)	(0.50)	(0.50)	(0.49)	(0.43)	(0.50)	(0.50)	(0.49)	(0.43)	(0.50)	(0.50)	(0.49)	(0.42)
Hispanic	0.18	0.24	0.20	0.15	0.07	0.24	0.20	0.15	0.08	0.23	0.20	0.15	0.08
	(0.38)	(0.43)	(0.40)	(0.35)	(0.26)	(0.43)	(0.40)	(0.35)	(0.28)	(0.42)	(0.40)	(0.35)	(0.27)
White	0.26	0.14	0.21	0.34	0.43	0.15	0.22	0.34	0.41	0.13	0.21	0.34	0.45
	(0.44)	(0.35)	(0.41)	(0.47)	(0.50)	(0.35)	(0.41)	(0.47)	(0.49)	(0.34)	(0.41)	(0.48)	(0.50)
FRPL	0.71	0.88	0.80	0.60	0.35	0.89	0.80	0.60	0.37	0.89	0.80	0.60	0.36
	(0.45)	(0.32)	(0.40)	(0.49)	(0.48)	(0.32)	(0.40)	(0.49)	(0.48)	(0.31)	(0.40)	(0.49)	(0.48)
English language learner	0.12	0.28	0.14	0.05	0.02	0.28	0.14	0.05	0.02	0.27	0.15	0.05	0.02
	(0.33)	(0.45)	(0.35)	(0.21)	(0.13)	(0.45)	(0.35)	(0.21)	(0.12)	(0.44)	(0.35)	(0.22)	(0.14)
Students with disabilities	0.05	0.08	0.05	0.03	0.02	0.08	0.05	0.03	0.02	0.08	0.05	0.03	0.02
	(0.21)	(0.27)	(0.22)	(0.17)	(0.14)	(0.26)	(0.22)	(0.17)	(0.15)	(0.27)	(0.22)	(0.17)	(0.14)
Observations	154475	25688	59326	45359	14585	25998	60811	46159	14783	22611	65117	47295	12769

Notes: Standard deviations in parentheses. BB means Below Basic, Prof means Proficiency, and Adv means Advanced.

Table 3. Regression discontinuity results with bandwidth of 2, math

	Benchmark A			Benchmark B			Benchmark C		
	BB/B	B/P	P/A	BB/B	B/P	P/A	BB/B	B/P	P/A
T	-0.013	-0.005	0.010	-0.020	-0.023	0.054*	0.033	-0.011	0.024
	(0.019)	(0.018)	(0.025)	(0.020)	(0.017)	(0.022)	(0.019)	(0.016)	(0.021)
T x waiver	0.013	-0.010	0.029	-0.031*	-0.010	0.034*	-0.040*	-0.024	-0.018
	(0.016)	(0.014)	(0.021)	(0.016)	(0.013)	(0.018)	(0.016)	(0.014)	(0.015)
Slope left	0.047***	0.060***	0.048***	0.086***	0.090***	0.038**	0.063***	0.081***	0.076***
	(0.011)	(0.009)	(0.013)	(0.012)	(0.010)	(0.012)	(0.012)	(0.009)	(0.011)
Slope right	0.055***	0.066***	0.080***	0.065***	0.072***	0.070***	0.066***	0.074***	0.090***
	(0.005)	(0.005)	(0.009)	(0.005)	(0.005)	(0.007)	(0.005)	(0.005)	(0.006)
Black	-0.058***	-0.062***	-0.075***	-0.040***	-0.051***	-0.094***	-0.049***	-0.043***	-0.084***
	(0.010)	(0.010)	(0.015)	(0.010)	(0.010)	(0.012)	(0.011)	(0.009)	(0.011)
Hispanic	0.021	0.026*	0.001	0.028*	0.016	-0.030*	0.012	0.011	-0.020
	(0.013)	(0.013)	(0.019)	(0.013)	(0.012)	(0.015)	(0.014)	(0.011)	(0.014)
Other race	0.096***	0.108***	0.099***	0.094***	0.089***	0.088***	0.051	0.053**	0.074***
	(0.024)	(0.018)	(0.020)	(0.028)	(0.019)	(0.018)	(0.028)	(0.018)	(0.016)
FRPL	-0.061***	-0.061***	-0.063***	-0.066***	-0.066***	-0.041***	-0.061***	-0.051***	-0.047***
	(0.008)	(0.008)	(0.011)	(0.010)	(0.008)	(0.009)	(0.009)	(0.007)	(0.009)
ELL	-0.021	0.019	0.032	0.012	0.019	0.077*	0.036	0.002	0.007
	(0.024)	(0.027)	(0.040)	(0.025)	(0.026)	(0.034)	(0.023)	(0.025)	(0.036)
SWD	-0.112***	-0.142***	-0.075*	-0.096***	-0.089***	-0.077*	-0.083***	-0.058*	-0.026
	(0.020)	(0.024)	(0.034)	(0.021)	(0.024)	(0.034)	(0.021)	(0.023)	(0.027)
Prior test score	0.548***	0.554***	0.477***	0.468***	0.487***	0.464***	0.433***	0.420***	0.422***
	(0.007)	(0.007)	(0.010)	(0.007)	(0.007)	(0.008)	(0.008)	(0.007)	(0.008)
Observations	31054	26982	12247	25335	27476	15781	22456	25007	17208
R^2	0.414	0.455	0.398	0.397	0.455	0.452	0.406	0.420	0.457

Notes: *p<0.05; **p<0.01, ***p<0.001. “T” represents the discontinuity in outcomes for students who received the higher of the two labels. Standard errors in parentheses are clustered at the school by grade by year level. All models condition on grade by year fixed effects, and school fixed effects. B/BB represents the Below Basic/Basic threshold, B/P represents the Basic/Proficient threshold, and P/A represents the Proficient/Advanced threshold.

Table 4. Regression discontinuity results with bandwidth of 2, reading

	Benchmark A			Benchmark B			Benchmark C		
	BB/B	B/P	P/A	BB/B	B/P	P/A	BB/B	B/P	P/A
T	0.009	-0.004	0.019	0.002	-0.015	-0.010	0.032	-0.017	0.003
	(0.026)	(0.017)	(0.021)	(0.024)	(0.016)	(0.021)	(0.026)	(0.016)	(0.020)
T x waiver	-0.021	-0.009	-0.012	0.001	0.014	-0.010	-0.046*	-0.008	0.026
	(0.018)	(0.012)	(0.018)	(0.018)	(0.013)	(0.017)	(0.021)	(0.013)	(0.017)
Slope left	0.019	0.057***	0.058***	0.033*	0.044***	0.075***	0.027	0.053***	0.070***
	(0.015)	(0.010)	(0.012)	(0.014)	(0.009)	(0.011)	(0.016)	(0.009)	(0.011)
Slope right	0.048***	0.046***	0.058***	0.048***	0.061***	0.063***	0.057***	0.069***	0.070***
	(0.006)	(0.005)	(0.008)	(0.006)	(0.005)	(0.008)	(0.006)	(0.005)	(0.008)
Black	-0.073***	-0.072***	-0.097***	-0.074***	-0.079***	-0.091***	-0.059***	-0.056***	-0.099***
	(0.013)	(0.009)	(0.013)	(0.012)	(0.009)	(0.012)	(0.014)	(0.008)	(0.012)
Hispanic	0.012	0.008	-0.017	0.001	-0.024*	0.004	0.006	-0.000	-0.034*
	(0.015)	(0.011)	(0.017)	(0.016)	(0.011)	(0.016)	(0.018)	(0.011)	(0.017)
Other race	-0.020	0.042*	0.058**	-0.030	0.014	0.062***	-0.012	0.052**	0.034
	(0.032)	(0.019)	(0.018)	(0.032)	(0.018)	(0.019)	(0.032)	(0.017)	(0.017)
FRPL	-0.083***	-0.055***	-0.065***	-0.058***	-0.055***	-0.066***	-0.050***	-0.055***	-0.060***
	(0.011)	(0.007)	(0.010)	(0.011)	(0.007)	(0.010)	(0.012)	(0.007)	(0.010)
ELL	-0.041	0.009	0.011	-0.032	0.009	0.043	-0.077**	0.005	0.019
	(0.028)	(0.025)	(0.061)	(0.023)	(0.025)	(0.048)	(0.026)	(0.026)	(0.055)
SWD	-0.136***	-0.101***	-0.058	-0.124***	-0.116***	-0.021	-0.146***	-0.095***	-0.015
	(0.024)	(0.022)	(0.033)	(0.023)	(0.020)	(0.031)	(0.024)	(0.021)	(0.030)
Prior test score	0.543***	0.547***	0.494***	0.508***	0.526***	0.508***	0.500***	0.517***	0.500***
	(0.008)	(0.006)	(0.009)	(0.008)	(0.006)	(0.008)	(0.009)	(0.006)	(0.008)
Observations	15529	23606	14994	15675	24197	15959	13429	24955	15485
R^2	0.391	0.433	0.418	0.375	0.436	0.445	0.430	0.445	0.449

Notes: *p<0.05; **p<0.01, ***p<0.001. "T" represents the discontinuity in outcomes for students who received the higher of the two labels. Standard errors in parentheses are clustered at the school by grade by year level. All models condition on grade by year fixed effects, and school fixed effects. B/BB represents the Below Basic/Basic threshold, B/P represents the Basic/Proficient threshold, and P/A represents the Proficient/Advanced threshold.

Table 5. Estimated discontinuities at label thresholds during NCLB and waiver period, math

		Below Basic/Basic					Basic/Proficient					Proficient/Advanced				
	Bandwidth (number of questions) →	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
Benchmark A	T	-0.013	-0.006	-0.004	-0.004	0.004	-0.005	0.011	0.011	0.020	0.024*	0.010	0.006	0.020	0.015	0.015
		(0.019)	(0.014)	(0.012)	(0.011)	(0.011)	(0.018)	(0.013)	(0.011)	(0.010)	(0.010)	(0.025)	(0.018)	(0.015)	(0.014)	(0.014)
	T x	0.013	0.008	-0.005	-0.004	-0.004	-0.010	-0.013	-0.027*	-0.032**	-0.030*	0.029	-0.004	-0.018	-0.023	-0.023
	waiver	(0.016)	(0.014)	(0.013)	(0.013)	(0.013)	(0.014)	(0.013)	(0.012)	(0.012)	(0.012)	(0.021)	(0.018)	(0.017)	(0.016)	(0.016)
	Obs	31054	41705	49266	53929	56255	26982	37387	47021	53380	56529	12247	17322	22809	27932	30933
	R^2	0.414	0.436	0.458	0.477	0.488	0.455	0.488	0.522	0.540	0.548	0.398	0.441	0.479	0.512	0.532
Benchmark B	T	-0.020	0.004	0.007	0.020	0.023*	-0.023	0.010	0.019	0.026**	0.026**	0.054*	0.020	0.023	0.012	0.012
		(0.020)	(0.015)	(0.013)	(0.012)	(0.011)	(0.017)	(0.013)	(0.011)	(0.010)	(0.010)	(0.022)	(0.017)	(0.015)	(0.014)	(0.013)
	T x	-0.031*	-0.032*	-0.032*	-0.033**	-0.032*	-0.010	-0.027*	-0.035**	-0.034**	-0.033**	0.034*	0.030	0.028	0.019	0.012
	waiver	(0.016)	(0.014)	(0.013)	(0.013)	(0.013)	(0.013)	(0.012)	(0.011)	(0.011)	(0.011)	(0.018)	(0.016)	(0.014)	(0.014)	(0.014)
	Obs	25335	34504	42531	48015	51369	27476	37873	46913	53427	56835	15781	22170	28444	33488	36134
	R^2	0.397	0.426	0.457	0.481	0.497	0.455	0.487	0.518	0.538	0.546	0.452	0.489	0.521	0.546	0.562
Benchmark C	T	0.033	0.039**	0.043***	0.047***	0.044***	-0.011	-0.012	-0.015	-0.008	-0.002	0.024	-0.002	0.001	0.008	0.012
		(0.019)	(0.015)	(0.012)	(0.011)	(0.011)	(0.016)	(0.012)	(0.010)	(0.009)	(0.009)	(0.021)	(0.015)	(0.013)	(0.012)	(0.012)
	T x	-0.040*	-0.056***	-0.059***	-0.064***	-0.066***	-0.024	-0.023	-0.028**	-0.031**	-0.030**	-0.018	0.004	0.016	0.018	0.015
	waiver	(0.016)	(0.014)	(0.013)	(0.013)	(0.013)	(0.014)	(0.012)	(0.011)	(0.011)	(0.011)	(0.015)	(0.014)	(0.013)	(0.013)	(0.013)
	Obs	22456	30638	38327	44149	47640	25007	34500	43508	50579	54713	17208	23687	30002	35616	38519
	R^2	0.406	0.436	0.474	0.493	0.503	0.420	0.459	0.505	0.538	0.557	0.457	0.498	0.535	0.566	0.584

Notes: *p<0.05; **p<0.01, ***p<0.001. “T” represents the discontinuity in outcomes for students who received the higher of the two labels. Standard errors in parentheses are clustered at the school by grade by year level. All models condition on student covariates (race, ethnicity, FRPL, ELL, and disability status), school covariates (log enrollment, prior year’s percent proficient in math, percent of school that is black, Hispanic, white, FRPL, ELL, and has a disability), grade by year fixed effects, and school fixed effects.

Table 6. Estimated discontinuities at label thresholds during NCLB and waiver period, reading

		Below Basic/Basic					Basic/Proficient					Proficient/Advanced				
	Bandwidth (number of questions) →	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
Benchmark A	T	0.009	-0.005	-0.001	0.009	0.016	-0.004	0.004	0.011	0.014	0.017*	0.019	0.007	0.005	0.011	0.010
		(0.026)	(0.018)	(0.016)	(0.014)	(0.013)	(0.017)	(0.012)	(0.010)	(0.009)	(0.008)	(0.021)	(0.016)	(0.013)	(0.012)	(0.011)
	T x waiver	-0.021	-0.020	-0.009	-0.020	-0.026*	-0.009	-0.014	-0.009	-0.007	-0.009	-0.012	-0.005	-0.006	0.000	0.001
		(0.018)	(0.015)	(0.014)	(0.013)	(0.012)	(0.012)	(0.011)	(0.010)	(0.009)	(0.009)	(0.018)	(0.016)	(0.015)	(0.015)	(0.014)
	Obs	15529	21600	27280	32748	38032	23606	32528	40921	48993	55532	14994	20649	25758	30544	35145
	R^2	0.391	0.413	0.431	0.454	0.479	0.433	0.471	0.512	0.551	0.582	0.418	0.464	0.501	0.529	0.552
Benchmark B	T	0.002	0.002	0.005	0.008	0.024*	-0.015	-0.021	-0.015	-0.016	-0.018*	-0.010	0.002	0.011	0.021	0.021
		(0.024)	(0.017)	(0.015)	(0.013)	(0.012)	(0.016)	(0.012)	(0.010)	(0.009)	(0.008)	(0.021)	(0.015)	(0.013)	(0.012)	(0.011)
	T x waiver	0.001	-0.004	-0.011	-0.027*	-0.030*	0.014	0.009	0.006	0.006	0.006	-0.010	-0.022	-0.018	-0.027	-0.031*
		(0.018)	(0.015)	(0.014)	(0.013)	(0.012)	(0.013)	(0.011)	(0.010)	(0.010)	(0.009)	(0.017)	(0.015)	(0.015)	(0.014)	(0.014)
	Obs	15675	21689	27444	32597	37199	24197	33500	42403	50246	56379	15959	21993	27483	32719	37312
	R^2	0.375	0.406	0.431	0.458	0.476	0.436	0.479	0.520	0.559	0.591	0.445	0.481	0.511	0.542	0.563
Benchmark C	T	0.032	0.022	0.019	0.027*	0.030*	-0.017	-0.012	-0.010	-0.004	-0.005	0.003	-0.004	0.009	0.011	0.017
		(0.026)	(0.018)	(0.015)	(0.013)	(0.013)	(0.016)	(0.011)	(0.010)	(0.009)	(0.008)	(0.020)	(0.015)	(0.013)	(0.012)	(0.011)
	T x waiver	-0.046*	-0.041*	-0.045**	-0.045**	-0.053***	-0.008	-0.014	-0.013	-0.018	-0.019*	0.026	0.035*	0.028	0.018	0.017
		(0.021)	(0.018)	(0.016)	(0.015)	(0.015)	(0.013)	(0.011)	(0.010)	(0.009)	(0.009)	(0.017)	(0.016)	(0.015)	(0.015)	(0.014)
	Obs	13429	18843	24125	29452	34083	24955	34425	43457	51542	57801	15485	21066	26234	31246	35756
	R^2	0.430	0.447	0.468	0.493	0.508	0.445	0.484	0.533	0.573	0.604	0.449	0.486	0.519	0.548	0.566

Notes: *p<0.05; **p<0.01, ***p<0.001. “T” represents the discontinuity in outcomes for students who received the higher of the two labels. Standard errors in parentheses are clustered at the school by grade by year level. All models condition on student covariates (race, ethnicity, FRPL, ELL, and disability status), school covariates (log enrollment, prior year’s percent proficient in reading, percent of school that is black, Hispanic, white, FRPL, ELL, and has a disability), grade by year fixed effects, and school fixed effects.

Figures

Figure 3. Percent of students assigned each performance label during NCLB and waiver, math

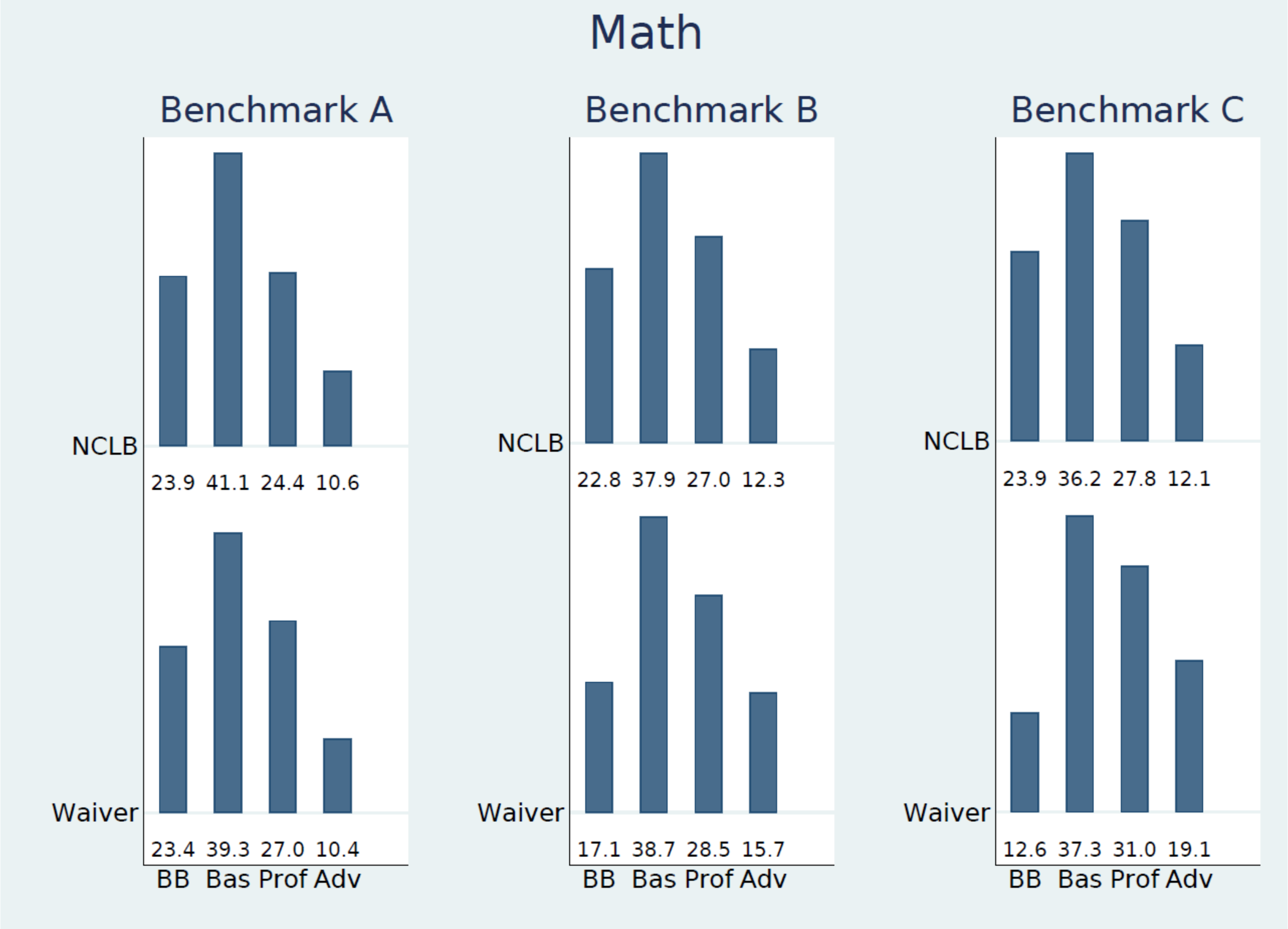


Figure 4. Percent of students assigned each performance label during NCLB and waiver, reading

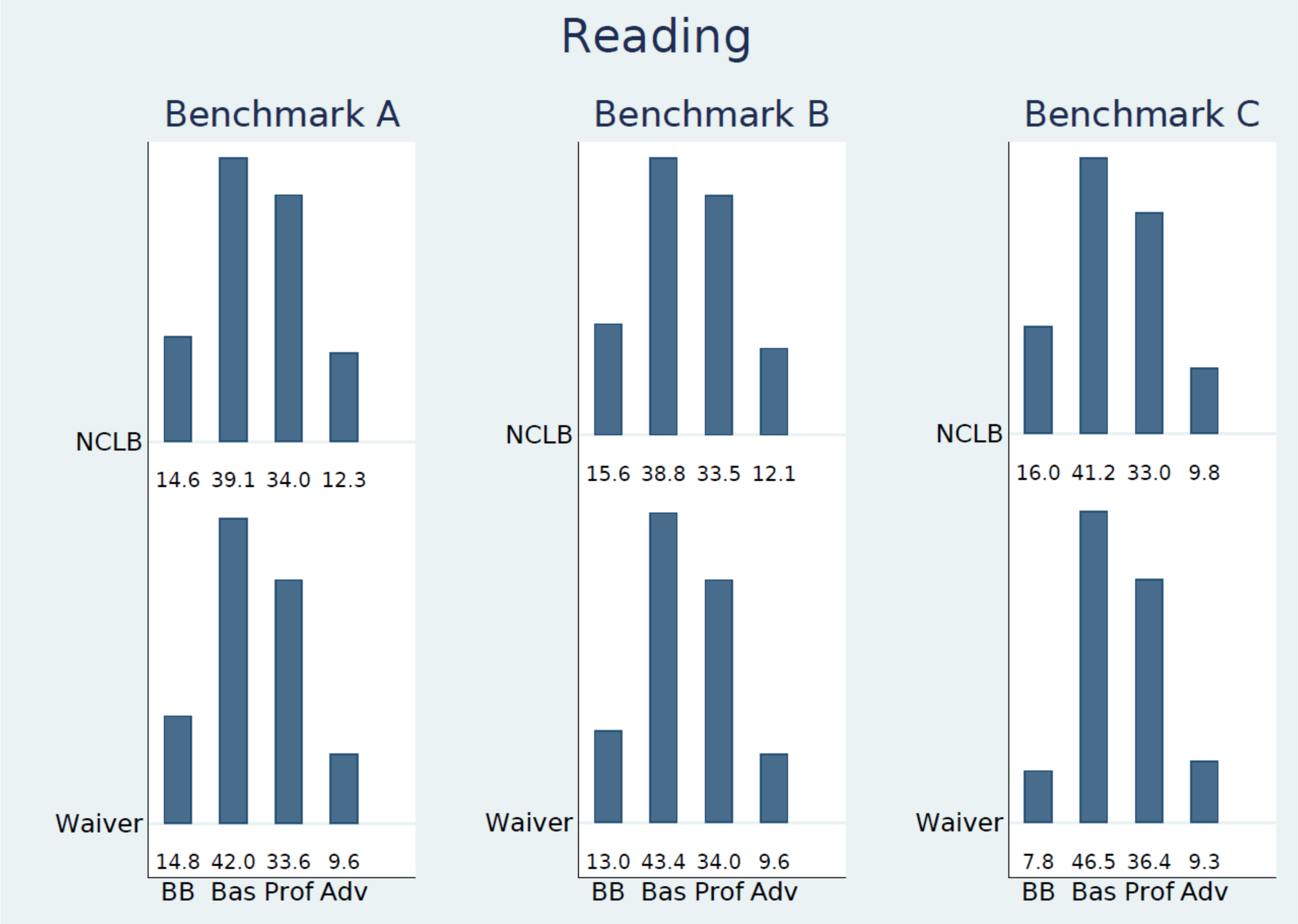


Figure 5. Estimated discontinuities at label thresholds during NCLB and waiver period, math

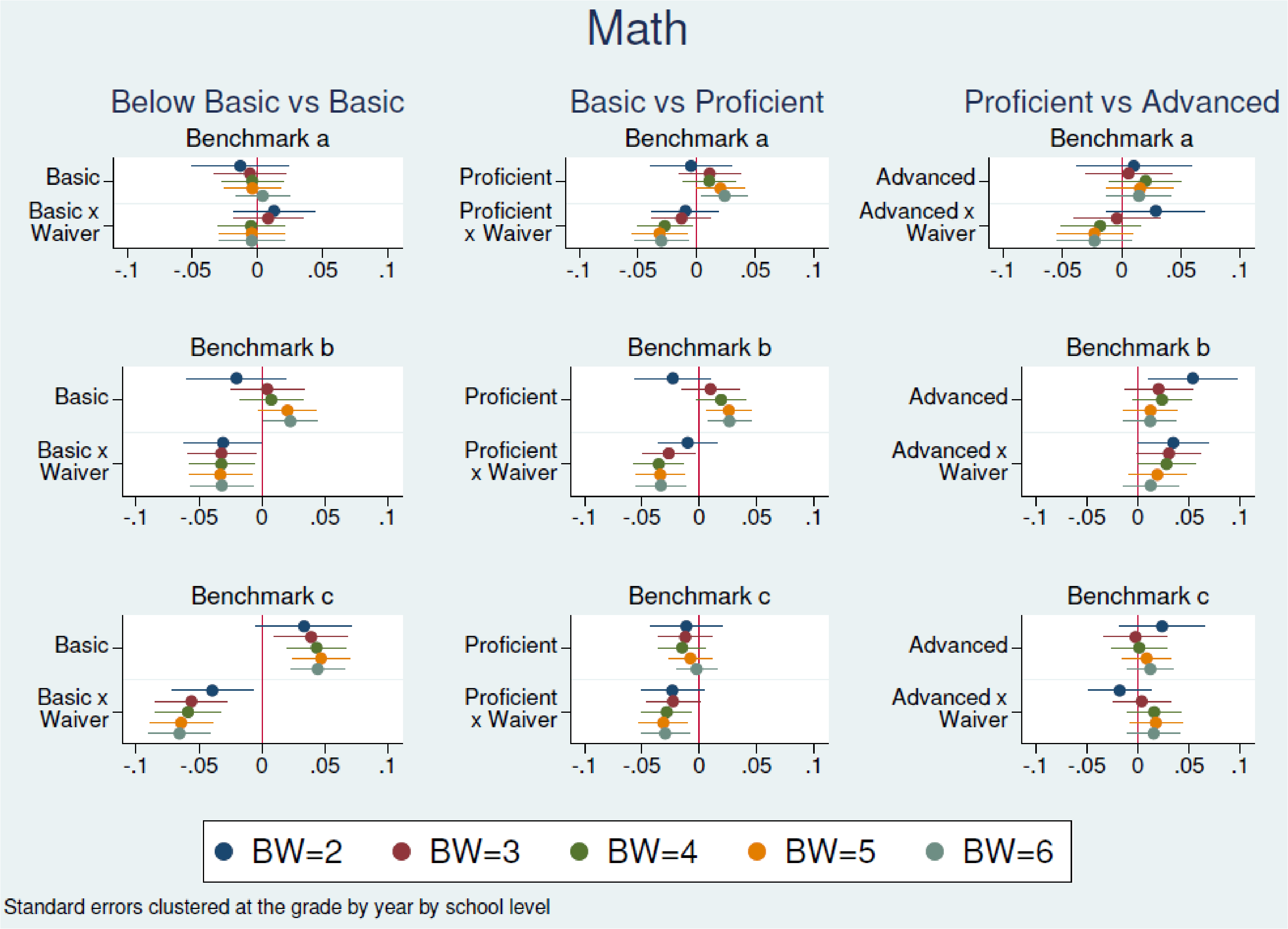


Figure 6. Estimated discontinuities at label thresholds during NCLB and waiver period, elementary schools math

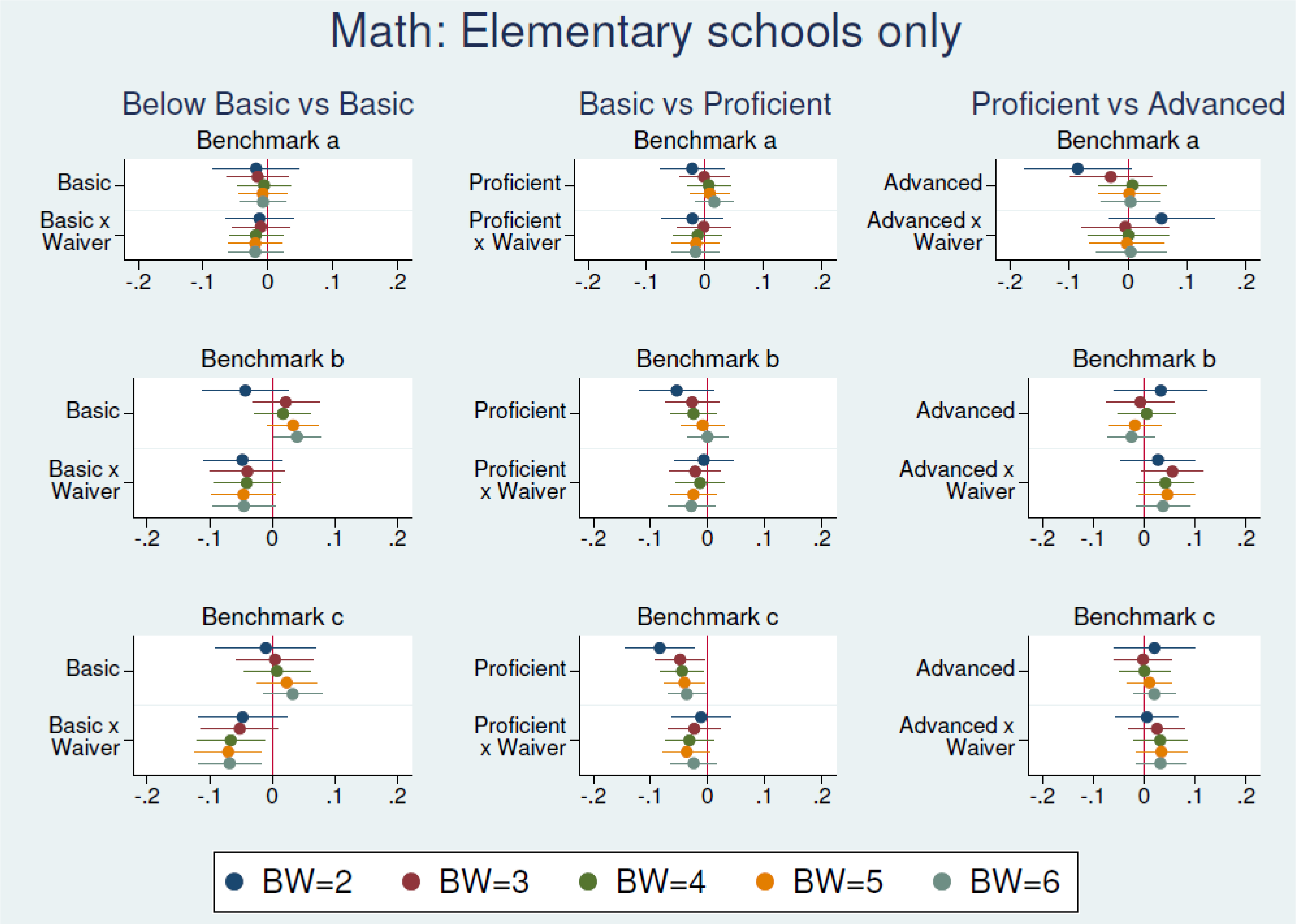


Figure 7. Estimated discontinuities at label thresholds during NCLB and waiver period, middle schools math

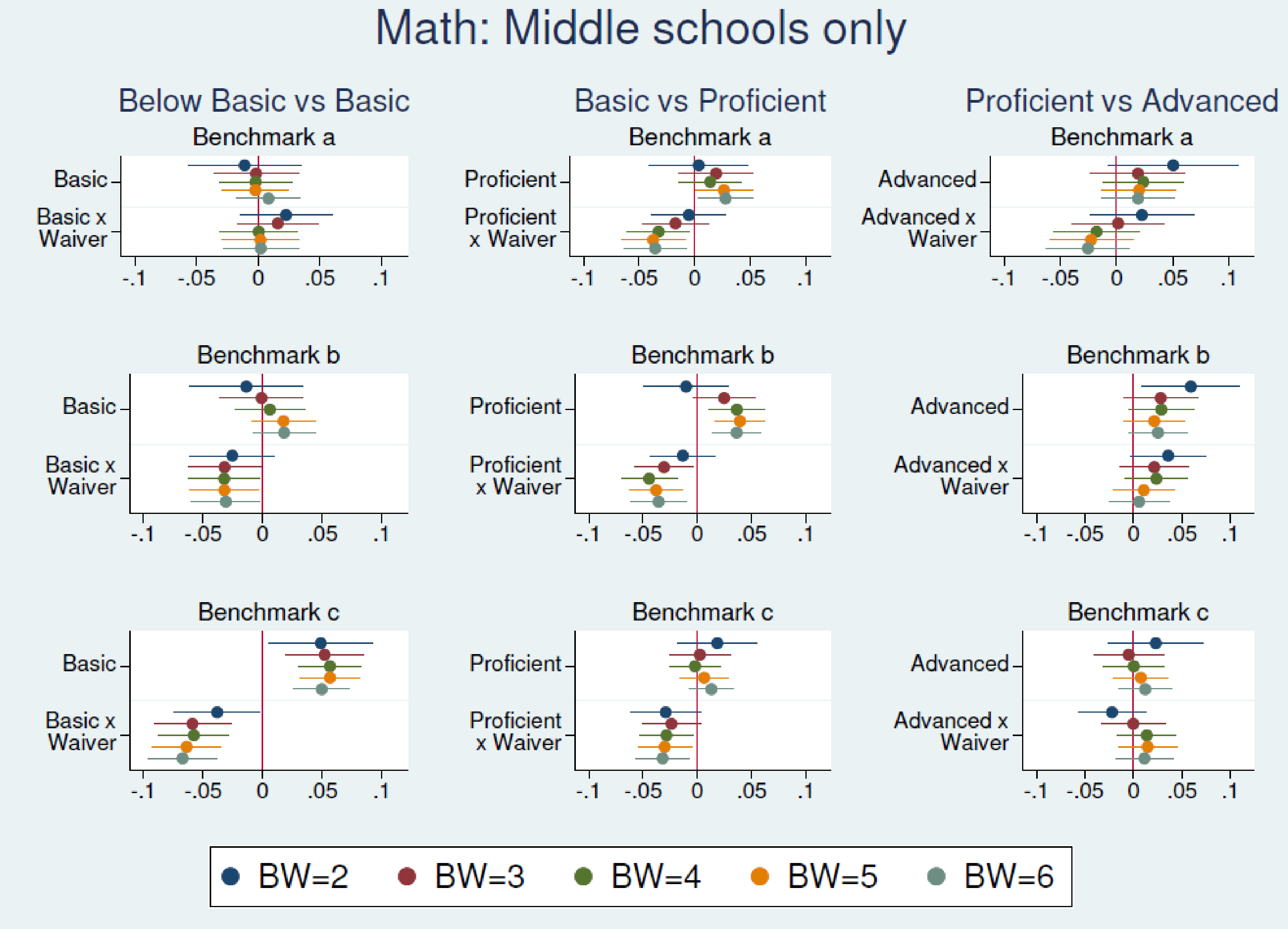


Figure 8. Estimated discontinuities at label thresholds during NCLB and waiver period, reading

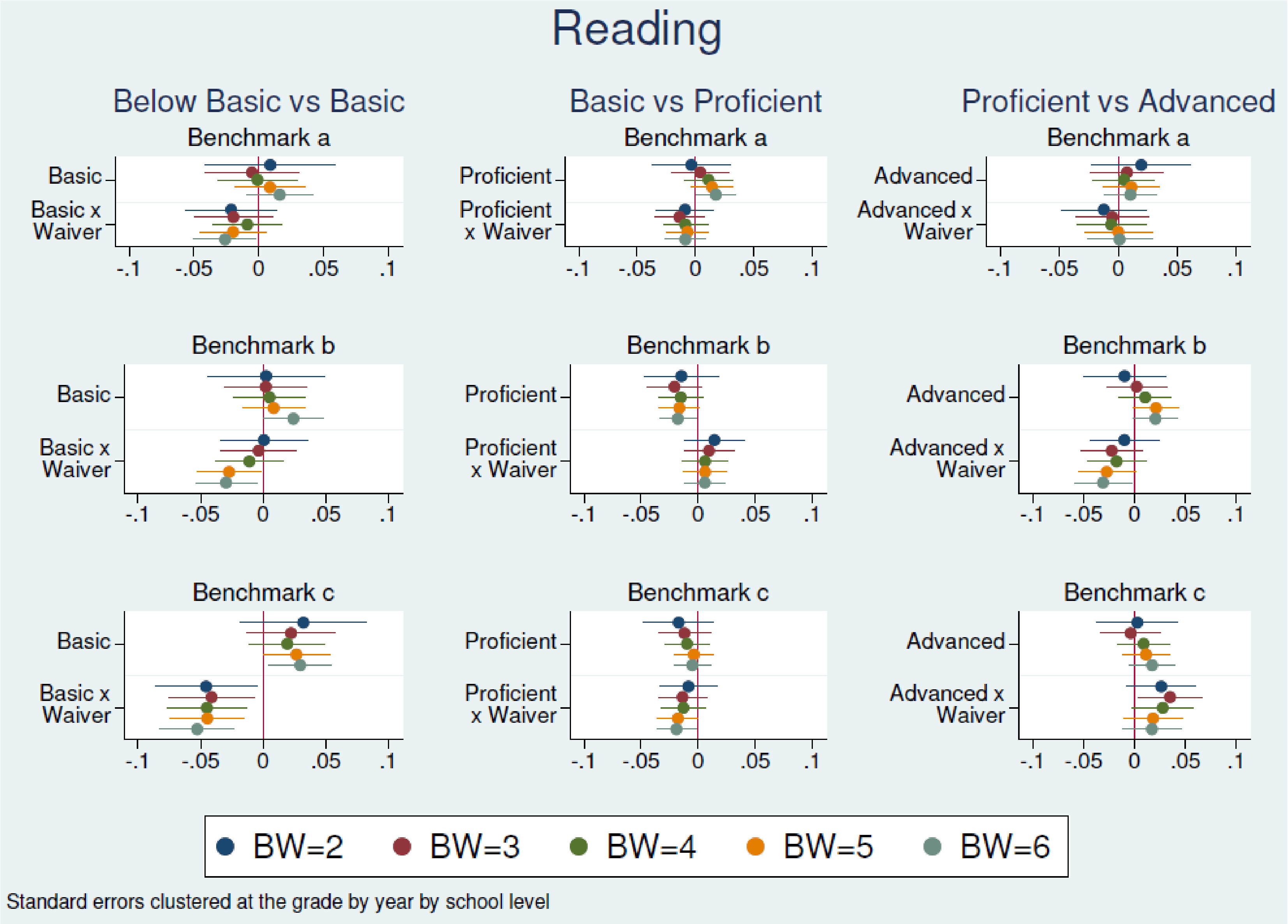


Figure 9. Estimated discontinuities at label thresholds during NCLB and waiver period, elementary schools reading

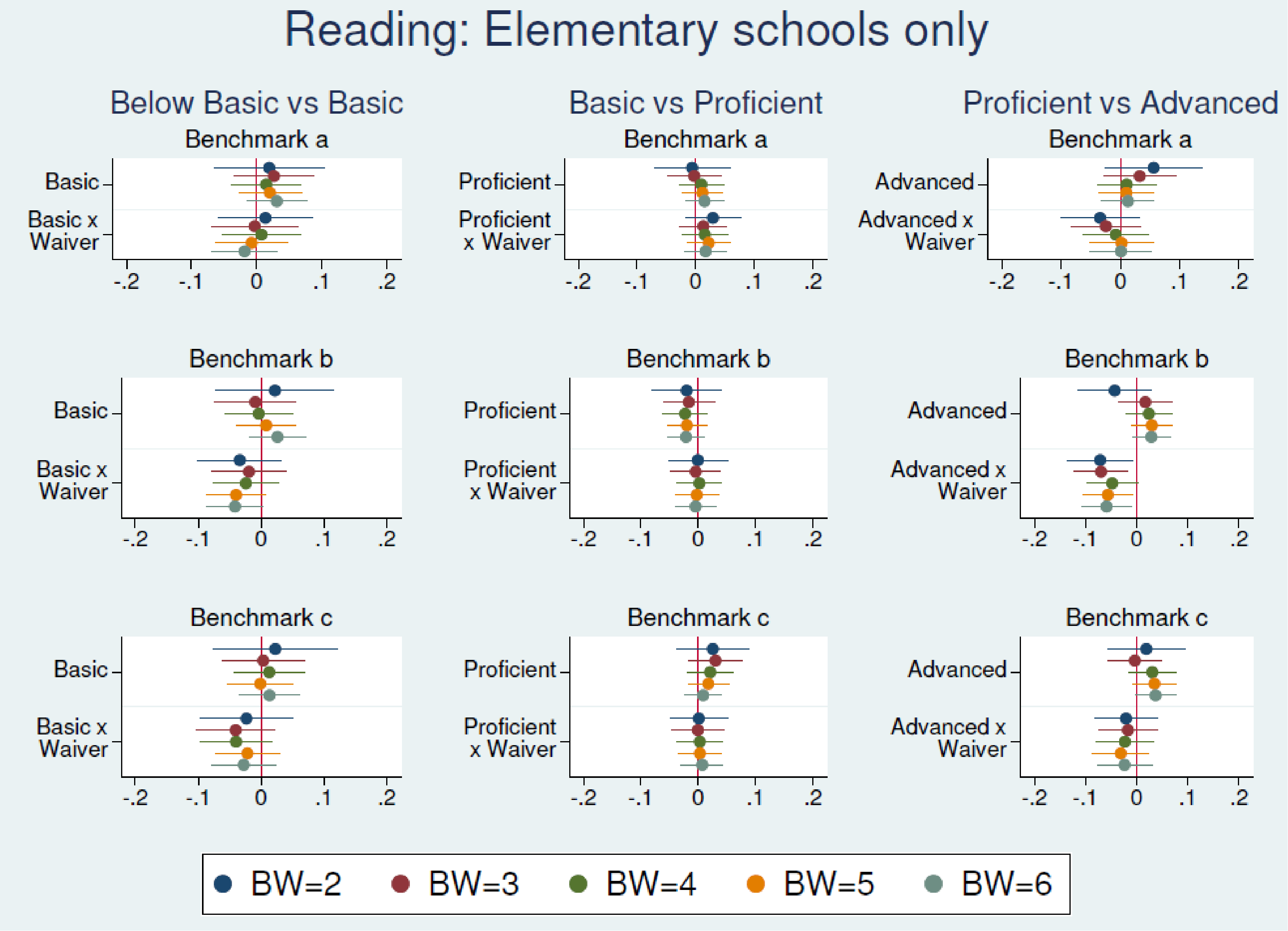
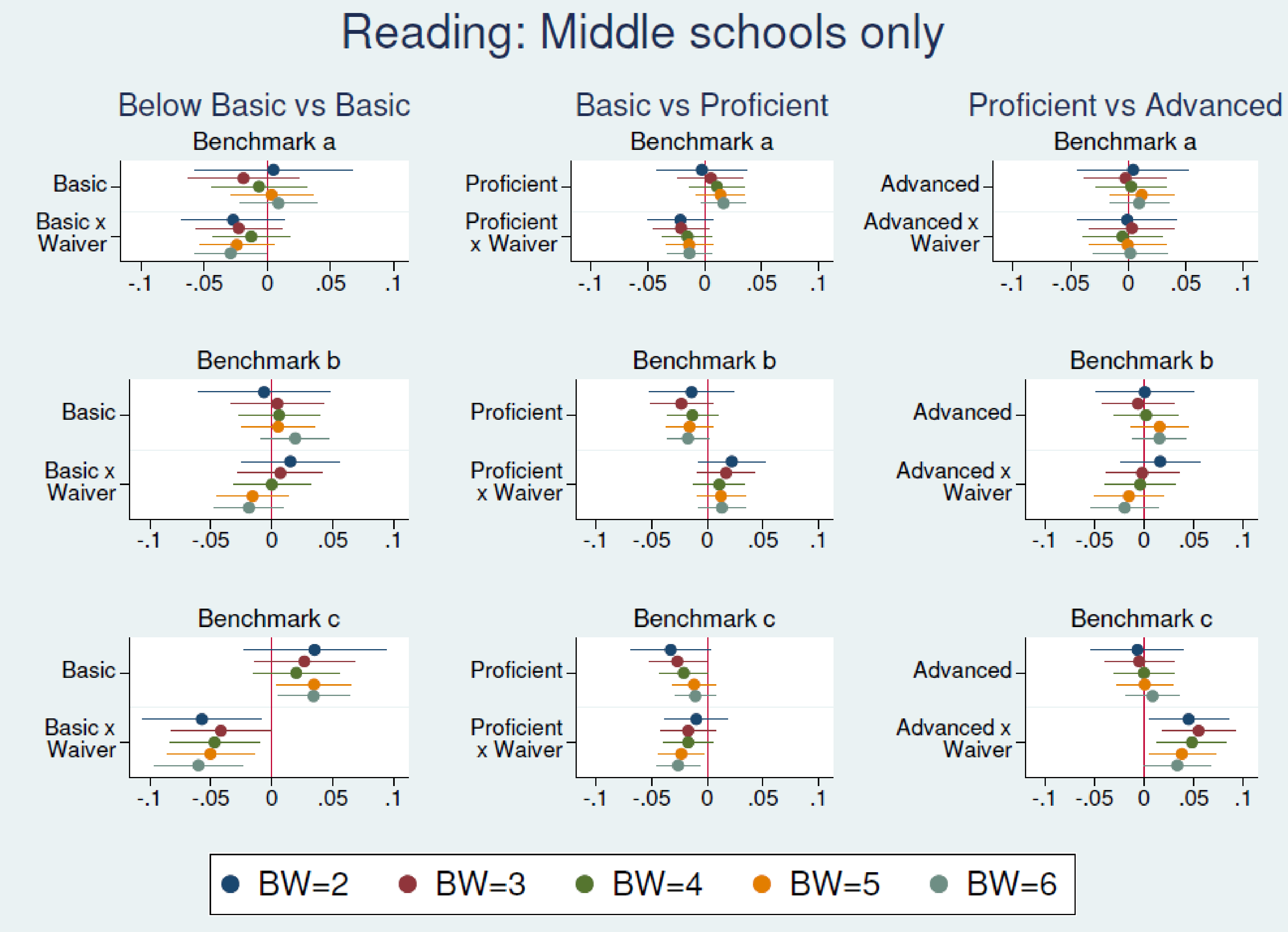


Figure 10. Estimated discontinuities at label thresholds during NCLB and waiver period, middle schools reading



Chapter 3: Alternative Explanations for the Reduced Focus on Bubble Students

The Chapter 2 results provide evidence that schools in Greenfield County Public Schools (GCPS) focused on bubble students during No Child Left Behind (NCLB) but were no longer doing so (at least in the same way) under the waiver accountability system. Those analyses show that lower-performing students were harmed during NCLB and benefitted after the waiver was implemented (2012-13 to 2013-14). One explanation for these results is that the onset of the waiver regime shifted the incentives facing schools, causing them to pay more attention to lower-achieving students. The introduction of the waiver is not the only explanation, however, and this chapter investigates three alternate explanations for the Chapter 2 results. After a brief introduction to the three hypotheses, I investigate each one in turn by (a) describing the data used and models estimated, (b) presenting the results, and (c) discussing the findings before moving onto the next alternative.

One alternate explanation is that the adoption of a new educator evaluation system, which occurred in 2011-12 (the year before the waiver was implemented), changed incentives so administrators and teachers lessened their focus on bubble students. Under the old evaluation system, educators were rated primarily based on observation scores from their supervisors. Under the new evaluation system, 35% of principal and teacher evaluation scores were comprised of value-add measures from their students' state test scores. The state evaluated teachers using these value-add scores in order to "maintain a focus on advancing each child against the current baseline results" (ESEA Flexibility Request, 2012, p. 34). These growth metrics would theoretically reduce the pressure for teachers to focus on proficiency rates (and bubble students). No important stakes were attached to the new evaluations for current teachers, however, which may limit how much these new incentives would induce teachers to change their behavior. On the other hand, the new incentives could have shifted behavior if teachers believed that higher stakes would be introduced over time or if they simply wanted to look good when evaluated.

A second alternate explanation for the increased focus on lower-performing students during the waiver is that schools realized that it did not make sense to draw distinctions between students who lay just to the left and just to the right of the Proficient threshold (or any threshold). Instead, schools may have changed the way they determined who was a bubble student. Under this hypothesis schools continued to practice triage, they simply changed the way they determined their bubble students beginning in 2012-13. Instead of using the test-score labels investigated in Chapter 2, schools may have used a district-furnished label which specifically identified the bubble students. If schools used this label to identify bubble students, this would explain why the regression discontinuity analysis from the previous chapter no longer found evidence of triage in the later years of the sample.

Beginning in 2012-13, the district combined multiple pieces of student-level data to classify students based on whether they had low-, moderate-, or high-probabilities of meeting proficiency on the state test. Students in those groups were labeled “Multi-Year Plan,” “Priority,” or “Enrichment,” respectively, as seen in the *Identification of Target Students* document shown in Figure 2. This document directed schools to “prioritize Target students” and was shared with all elementary and middle school principals. These actions—dividing students into three groups based on their closeness to meeting proficiency and directing schools to focus on the middle group—represent district-endorsed triage. The district basically said to schools: the students labeled Priority are the bubble students, and you should focus on them. Comparing relative differences in outcomes for students in these groups before and after the waiver can also allay any lingering concerns from the previous chapter that schools would not have treated students differently based on their benchmark labels.

A third alternative is that additional funding provided by the district mitigated the negative effects of triage on low-achievers. The district research team identified 29 low-performing schools that could benefit from additional federal funds available in 2012-13 (the first year of the waiver). Each of the 29 school leaders turned in a Targeted Academic Intervention Proposal (TAIP), detailing how the school would use the money to restructure their school day to offer targeted interventions. The additional TAIP funds for this selected group of schools could explain why the Chapter 2 analysis fails to find the focus on bubble students during the waiver. Triage occurs when there are limited resources: schools prioritize students close to proficiency and then divert resources to them. If schools received additional money to offer interventions for bubble students, then they would not need to divert resources from non-bubble students.

Alternative 1: Adoption of a New Educator Evaluation System in 2011-12

This analysis investigates the hypothesis that the increased focus on low-performers found in Chapter 2 was because the new educator evaluation system included value-add growth metrics. This is investigated by exploiting the fact that the new evaluation system was implemented a year before the waiver took effect.

Data

This analysis uses the same testing data from Chapter 2, except that the 2012-13 and 2013-14 students are dropped from the sample. A binary variable *Evaluation* indicates which evaluation system was in effect during the school year. A value of zero represents the old evaluation system (2009-10 through 2010-11), which primarily used observation scores to rate educators, and a value of one represents the first year of the new educator evaluation system (2011-12), when value-add growth metrics comprised 35% of teacher and principal ratings.

Methods

The same regression models from Chapter 2 (Equation 1) are estimated on the test-score labels before and after the new evaluation system was introduced. That is, ordinary least squares (OLS) is used to calculate the discontinuity at each threshold (Basic, Proficient, and Advanced) on each benchmark, and an interaction term is included to quantify the difference in the discontinuity when the new evaluation system was implemented. The equation is as follows:

$$Y_{sijt} = \beta_0 + \beta_1 T_{sijt} + \beta_2 (T_{sijt}\, x\, Evaluation_t) + \beta_3\, num\_correct_{sijt} + \beta_4 (T_{sijt}\, x\, num\_correct_{sijt}) + \delta X_{it} + \gamma Z_{jt} + \eta_j + \theta_t + e_{ijt} \tag{3}$$

In these models, Y_{sijt} represents the standardized end-of-year state test score on the subject-area test *s* for student *i* in school *j*, in year *t*. T_{sijt} represents treatment (a binary variable indicating that the student received the higher of the two performance labels in subject *s*), $Evaluation_t$ is a binary variable indicating that the state was under the new evaluation system in year *t*, $num\_correct_{sijt}$ represents the distance from the cutoff score (students who score below the cut-off have negative values for this, students who score at or above the cut-off have zero or positive values), X_{it} is a vector of student controls (for student *i* in year *t*, including prior year state test score, race, FRPL, ELL, and disability status), Z_{jt} is a vector of time-varying school controls (for school *j* in year *t*, including log enrollment, prior year's percent proficient in subject *s,* and the percent of students in the school who are Black, Hispanic, White, FRPL, ELL, and have a disability), η_j represent school fixed effects, θ_t are grade-by-year fixed effects, and e_{ijt} is the error term. Standard errors are clustered at the school by grade by year level. As with the previous chapter, these models are run across varying bandwidths of the number of benchmark questions answered correctly, from two to six questions on either side of each threshold.

Because the models are the same from Chapter 2, I now look for a change associated with the adoption of the evaluation system in 2011-12 in the same way that I previously looked for a change associated with the waiver. If schools behaved more equitably in response to the waiver incentives but not the evaluation system, then the estimated discontinuities at each threshold should be close to zero for the "*T x Evaluation*" coefficients. This would mean that there was no change in school behavior based on the adoption of the evaluation system prior to the waiver. On the other hand, if schools did respond to the incentives of the evaluation system by paying more attention to the lower-performing students, then the coefficient on the interaction would be negative[9].

Results

Because the first hypothesis is tested using the same methods from Chapter 2, the results are presented similarly. The estimated discontinuities at each label threshold before and after the implementation of the new evaluation system are shown in Table 7 (math) and Table 8 (reading), with those estimates graphed in Figures 11 and 12, respectively.

[9] One aspect of comparing 2011-12 outcomes with previous years is that it removes these observations from the previous chapter's NCLB estimates. As a preliminary step to see whether the Chapter 2 results depended on 2011-12 data being included in the NCLB period, I checked two variations to those analyses. These involved (a) recoding 2011-12 to be included as the waiver and (b) dropping 2011-12 altogether (so the NCLB era included 2009-10 and 2010-11 and the waiver era included 2012-13 and 2013-14). The math results for 2011-12 recoded as the waiver are shown in Appendix Figure B1 and with 2011-12 dropped altogether are shown in Appendix Figure B2. The same results for reading are shown in Appendix Figures B3 and B4, respectively. Corresponding results were very similar to those presented in Chapter 2, indicating that those results did not depend on including 2011-12 in the NCLB era.

Math. The inclusion of growth metrics in the educator evaluation system appears to have benefitted lower-performing students in math although not as broadly as reported in the previous chapter. Chapter 2 demonstrated that after the adoption of the waiver, there were consistent and significant negative estimates for the "*T x Waiver*" coefficient at both the Basic and Proficient thresholds and on both the second and third math benchmarks. The current results shown in Table 7 indicate that the new evaluation system led to a shift in focus only at the Basic threshold and only on math Benchmark C.

As with the previous chapter, the estimated discontinuities at the Basic threshold on the third math benchmark ("*T"*) are consistently positive (although the estimates in this chapter are statistically significant only at the higher bandwidths). This indicates that, prior to the evaluation policy change, students barely labeled Basic gained more than those barely labeled Below Basic. That is, bubble students gained more than low performers during NCLB. These estimates correspond to the Chapter 2 results and are similar in magnitude ($\beta \approx 0.040$). The differential effect of the Basic label after the adoption of the new evaluation system is consistently negative at the Basic threshold on the third benchmark (ranging from -0.010 to -0.045). Because the negative coefficients represent the differential effect for students who receive the higher label, these estimates indicate that students who barely scored Below Basic on Benchmark C gained about 0.04 standard deviations more in 2011-12 than similar students in previous years (estimates which are statistically significant except at the narrowest bandwidth). These gains counteract the negative effect of the Below Basic label found during NCLB. These results are similar to those from the previous chapter, although the estimated differential effect of the new evaluation system is slightly smaller in magnitude (the previous chapter estimated that the differential effect during the waiver was about 0.06 standard deviations).

Reading. As with math, the adoption of the new evaluation system appears to have benefitted lower-performing students near the Basic threshold on reading Benchmark C. The discontinuities for the time period prior to the policy change ("*T*") are positive and statistically significant at higher bandwidths (β= 0.030 to 0.040). These results are similar to those from Chapter 2 and provide more evidence that schools focused on Basic students over Below Basic students during NCLB. The estimated coefficients shown in Table 8 for "*T x Evaluation*" at the Basic threshold on Benchmark C are consistently negative (ranging from -0.010 to -0.065) but get larger in magnitude and are significant only as the bandwidth widens. This provides some evidence that students barely labeled Below Basic in reading gained more after the introduction of the new evaluation system. These negative differences in the "*Evaluation*" period indicate that the focus on bubble students was less evident in 2011-12. This suggests that educators responded to the change in incentives from the new evaluation system and shifted attention towards the lower-achieving students. The differential discontinuities for reading after that policy change are less consistent across bandwidths, however, than those from Chapter 2.

Discussion

The Chapter 2 results indicate that the waiver ameliorated the earlier negative effect of NCLB for students who barely score Below Basic on the third benchmark in both subjects. Those results are consistent across all bandwidths, which provides strong evidence that schools changed their behavior during the waiver period compared to NCLB. This first analysis investigated the hypothesis that the shift in focus to non-bubble students found during the waiver could be attributable to a change in educator incentives adopted the previous year. These results suggest that schools began to shift their attention towards lower-performing students in 2011-12, as would be predicted by the incentives created by including growth scores. This indicates that the changes to the evaluation system contributed to the gains found for lower-achieving students. Yet the results from the 2011-12 data show more variation in statistical significance across bandwidths than those from the previous chapter. Furthermore, the findings in this analysis are concentrated only at the Basic threshold on the third benchmark. This suggests that the new evaluation system was not the only factor in reducing the focus on bubble students, at least not in its first year. The evidence points toward the change in incentives created by the adoption of the waiver system or (possibly) the evaluation system as contributing to the changes found in Chapter 2.

The state intended for the waiver accountability system and the educator evaluation system to work in tandem (a) to increase overall student achievement and (b) to grow "every student, every year." From these results, it appears that the state was successful in shifting attention away from bubble students and down the test score distribution. The state changed incentives twice: once for evaluating educators and once for rating schools. It is possible that the Chapter 2 findings are the result of the second year of the evaluation system and not the introduction of the waiver. While I cannot disentangle the effects of these different policy changes, the explanation for both is the same: if incentives change, schools respond. The effects of these different policies are worth exploring further to see which incentives appear to cause shifting behavior. For instance, the new evaluation system applied only to new teachers. Further research might look in to whether the more equitable behavior is concentrated amongst new teachers or whether tenured teachers also responded to the new evaluation system.

Alternative 2: Schools Changed How They Defined Bubble Students during the Waiver

The second alternative examines the hypothesis that schools were still focusing on bubble students after the waiver was implemented but that they changed their definition of who the bubble students were. This could occur if schools realized that it did not make sense to treat students differently based on which side of a label threshold they scored. The hypothesis investigated here is that the district-supplied Priority label facilitated the change in how schools identified bubble students beginning in 2012-13, when GCPS created the more prescriptive labels and shared them with schools.

These prescriptive labels are valuable for several reasons to explore whether schools' behavior changed over time. First, the fact that the district provided these labels and directed schools to focus on Priority students makes it plausible that schools actually targeted these students. Because these labels identify the targeted group, this removes the guesswork from the analysis about whom the schools would consider bubble students. Second, the labels, the descriptions, and the color-coding of the *Identification of Target Students* document (Figure 2) reflect the three triage groups of too low, suitable, and too high, respectively. Third, these groups overlap the thresholds tested in Chapter 2, with students close to the Basic threshold labeled "Multi-Year Plan" (MYP), students close to the Proficiency threshold (i.e., the bubble students) labeled "Priority," and students close to the Advanced threshold labeled "Enrichment."

It is also plausible that even before receiving these prescriptive labels in 2012-13, schools identified bubble students not based on the Basic label (as investigated in Chapter 2) but instead in a way that is similar to the district's Priority label. That is, schools viewed students as "on the bubble" if they answered within a few questions above and below the Proficient label on the benchmark. This analysis also explores this hypothesis by using the district definitions to identify which students would have received the prescriptive labels in the NCLB era.

Data

This analysis uses the same student data as Chapter 2 (2009-10 through 2013-14) with several additional variables included.

Projected performance. Beginning in 2011-12, the state electronically shared with districts individual student projections in math and reading. These projections range from 0 to 100 and were calculated separately for math and reading based on each student's entire test history. The projection variables shared by the state included (a) the projected percentile for the student's end-of-year performance, (b) the probability that the student would score Advanced, (c) the probability that the student would score Proficient, and (d) the probability that the student would score Basic. In 2012-13, the district used the individual probabilities provided by the state to classify students as Advanced, Proficient, Basic, or Below Basic as follows (these performance labels represent each student's *state projection* and are separate from the benchmark labels assigned to students):

- If P(Advanced) ≥ 50%, the district labeled students "Advanced"
- If P(Proficient) ≥ 50% and P(Advanced) < 50%, the district labeled students "Proficient"
- If P(Basic) ≥ 50% and P(Proficient) < 50%, the district labeled students "Basic"
- If P(Basic) < 50%, the district labeled students "Below Basic"

Prescriptive labels. In 2012-13, the district research office used two different metrics to assign students prescriptive labels. In that year, each student's *current performance* and *state projection* placed them on a 5 x 4 matrix, as shown in the *Identification of Target Students* document in Figure 2. The test-score label from Benchmark B comprises the top row in Figure 2 and represents each student's *current performance*. As analyzed in the previous chapter, each student's second benchmark score was assigned a label of Advanced, Proficient, Basic, or Below Basic based on the number of questions answered correctly. The district used each student's label derived from the *state projection*, as described above, to comprise the left-hand column in Figure 2. Based on the combination of their current and projected performance, each individual was labeled as Priority 1, Priority 2, Enrichment, MYP, or No Label. Students who were missing a state projection (primarily third graders who had not yet taken a state test) had their prescriptive label determined solely by their benchmark score, as shown in the bottom row of Figure 2. Using the combination of their benchmark scores and state projections, I assigned each student in the dataset their prescriptive label of MYP, Priority[10], or Enrichment[11].

This analysis is intended to investigate whether schools changed the way they identified their bubble students during the waiver by using this district-provided Priority label. To examine whether this differed from how they behaved during NCLB, I assigned prescriptive labels to students in previous years based on what label they would have received if the district had

[10] As will be shown in the next chapter, school leaders did not distinguish between Priority 1 and Priority 2 students. I follow this convention and call these students collectively "Priority" students.

[11] The district did not provide schools with a list of students and their corresponding prescriptive label. Instead, they created a report in the online data warehouse called the *Virtual Data Wall Report* (shown in Appendix Figure B5), which schools could use to identify their Priority students. Because of this, it is unclear in some cases which prescriptive label a student would have received from their school. While the district provided guidelines to schools regarding these labels, they allowed school leaders some leeway in determining their Priority students. As noted at the bottom of the *Identification of Target Students* matrix (Figure 2), the district indicated that the students whose scores have them fall "within a few (3-4) items of [the] number correct cut score for Proficient level" should be included in the Priority group. This flexibility means that each individual school could determine whether their "bubble" would (a) include three questions on either side of the Proficiency line or (b) be broadened to include students within four question on either side. Appendix Figure B6 recreates the *Identification of Target Students* matrix to show the leeway provided to schools in defining these labels. The dotted lines in Appendix Figure B5 indicate that some of the labels are flexible and determined at the school level. Schools were given flexibility to include students within four question of proficiency only for these groups of students: (a) students labeled Proficient on both metrics, (b) students labeled Basic on both metrics, (c) students missing a projection and scoring Proficient on the benchmark, and (d) students missing a projection and scoring Basic on the benchmark. Because it is unclear which label students who were exactly four questions from proficiency would be assigned, I opted to leave their label blank. About 7% of the sample are in one of those four groups and answer exactly four questions above or below proficiency.

created the prescriptive labels earlier. Because the state supplied projections in 2011-12, I use the same methods just described to assign prescriptive labels to students in that year (even though the district did not use that information until 2012-13). For the 2009-10 and 2010-11 school years, when the state did not electronically share projected performance, I use the assignment rules from the bottom row of Figure 2 to classify students (which are for students who are missing projections).

To be clear, schools had access to the specific prescriptive labels of MYP, Priority, and Enrichment only in 2012-13 and 2013-14. These labels correspond with the students whom schools in earlier eras might have considered low, bubble, and high, respectively. This definition of bubble students—scoring some distance on either side of proficiency—is similar to strategies used in prior triage research (e.g., Ladd & Lauen, 2010; Springer, 2008b). The main difference is that rather than assigning more or less arbitrary bounds around the proficiency line, I use the district's definition of within three questions of proficiency on the benchmark. Another benefit of this district benchmark data is that Chapter 2 provides evidence that schools did use this assessment information, especially the third benchmark scores.

Figure 13 shows the percent of the sample labeled MYP, Priority, and Enrichment in math for each accountability era. During NCLB, the percent of students in each category is relatively consistent across math benchmarks: about 46% are MYP, 37% are Priority, and 17% are Enrichment. On the other hand, during the waiver, the percent of MYP students decreased from 45% on Benchmark A to 39% on Benchmark C with a corresponding increase in Enrichment students from 17% to 22%. Priority students in math during the waiver consistently represent about 39% of the sample. Figure 14 shows the same information for reading, when the labels are relatively consistent across benchmarks and eras. MYP students make up 45-48% of the reading sample, Priority students represent 31-34%, and Enrichment students comprise 20-23%.

Methods

In this analysis, the outcomes are compared across accountability eras based on whether students received (or would have received) the Priority, MYP, or Enrichment label. If schools targeted Priority students, as the district encouraged them to do beginning in 2012-13 year, then these students should gain significantly more than the other students. If schools distinguished between Below Basic and Basic students when they had those labels but then used the Priority label once it was supplied by the district, this could explain why Chapter 2 found that the Basic threshold no longer mattered during the waiver.

I use a difference-in-differences (DID) approach to examine this hypothesis empirically. This change from the local linear regression discontinuity (RD) method used in Chapter 2 reflects both (a) the hypothesis being investigated and (b) the district's classification of a group of students as priority. The underlying motivation for using local linear RD in the previous chapter is that schools would treat students differently based on if they were just above or just below a particular label threshold. While the previous chapter found discontinuities in average outcomes for students close to some thresholds, schools may have realized that it did not make a lot of sense to draw sharp distinctions between students with similar benchmark scores but who were clustered on opposite sides of a threshold. Schools may have instead viewed students close to the proficiency line as bubble students rather than distinguishing between Proficient and Basic students (something that is true at other thresholds as well). Other researchers have had to designate arbitrary upper and lower bounds around proficiency to identify the bubble students, but I am able to use the district-supplied label which effectively identified Priority students as deserving special treatment.

This DID method compares in one model the performance of the Priority students with those who are considered too low or too high to receive that label. Higher order polynomials of each assessment variable (benchmark score, prior year test score, projected performance) control for students' underlying achievement patterns which might explain why the outcomes of students labeled Priority differed from students labeled MYP or Enrichment. The binary variables representing the low- and high-achievers are included in the model to quantify the effect of the label beyond students' achievement histories. As with Chapter 2, the analyses are run separately for each of the three benchmarks.

I use the following DID model to estimate the effect of the label on student outcomes.

$$\begin{aligned} Y_{sijt} = {} & \beta_0 + \beta_1 \mathit{Multi\text{-}YearPlan}_{sijt} + \beta_2 \mathit{Enrichment}_{sijt} + \\ & \beta_3 (\mathit{Multi\text{-}YearPlan}_{sijt} \text{ x } \mathit{Waiver}) + \beta_4 (\mathit{Enrichment}_{sijt} \text{ x } \mathit{Waiver}) + \\ & \textstyle\sum_{j=1}^{k} \lambda_j Achievement_{sijt}^{j} + \delta X_{it} + \gamma Z_{jt} + \eta_j + \theta_t + e_{ijt} \end{aligned} \tag{4}$$

In this equation, Y_{sijt} represents the standardized state test score (for student i in subject s in school j in year t), $\mathit{Multi\text{-}YearPlan}_{sijt}$ represents a binary variable equaling one if the student was labeled as MYP, $\mathit{Enrichment}_{sijt}$ represents a binary variable equaling one if the student was labeled as Enrichment, *Waiver* is a binary variable equaling one if the year is 2012-13 or later, $\mathit{Achievement}_{sijt}$ represents various testing variables (up to cubic polynomials for each of prior year test score, projected percentile, benchmark raw score [centered at proficiency] and the interaction of each test score with one another)[12], X_{it} is a vector of student characteristics (race and ethnicity, FRPL, ELL status, and disability status), Z_{jt} is a vector of time-varying school controls (for school *j* in year *t*, including log enrollment, prior year's percent proficient in subject *s,* and the percent of students who are Black, Hispanic, White, FRPL, ELL, and have a disability), η_j represent school fixed effects, θ_t are grade-by-year fixed effects, and e_{ijt} is the error term. Standard errors are clustered at the school by grade by year level.

The targeted (Priority) students are the omitted group in this equation. Two coefficients of interest are β_1 and β_2, which represent the difference in outcome for low- and high-performers compared to bubble students during NCLB. Both of these coefficients would be negative if schools used these labels to engage in triage, which would support the hypothesis that schools used this alternate definition of bubble students during NCLB.

The other coefficients of interest are β_3 and β_4, which represent the differential outcome for low- and high-performers compared to bubble students during the waiver. Positive estimates for β_3 and β_4 would indicate that low- and high-achievers improved during the waiver. If schools shifted their definition of bubble students toward the group identified by the district Priority label in 2012-13, then β_3 and β_4 would be negative, implying that the low- and high-performers were

[12] A number of students were missing achievement data, especially prior year test score. Because 3rd graders take the state test for the first time at the end of the school year, they would be excluded from this analysis due to missing prior test scores. Because 3rd graders are part of the group that counts twice towards the accountability system during the waiver, it was important to include them in the analysis. To include students in the analysis who were missing some achievement data, all students with missing test variables were assigned a score of zero on that variable, and the test score variables from Equation 4 were interacted with a variable indicating each student's pattern of missingness for the achievement variables. This allows for each student to have their outcome predicted by using all of the test score information available while minimizing the number of students excluded from the analyses.

harmed by this shift in attention to Priority students. Due to the benchmark analyses run so far, the strongest evidence is expected on the third benchmark[13].

Results

The results from this model are shown in Table 9 for both math and reading. In math, there is evidence of triage based on the third benchmark during NCLB using this alternate definition of bubble students. On Benchmark C, both low- and high-performers gain significantly less than bubble students in these years. The estimate for low-performers (β= -0.031, $p<0.001$) is similar in magnitude to the discontinuities found in Chapter 2 for Below Basic students close to the Basic threshold. Furthermore, the results from this model indicate that the low-performing students (those labeled MYP by the district) benefitted during the waiver. The estimate for *MYP x Waiver* on the third benchmark is 0.043 ($p>0.001$), indicating that these students gained significantly more during the waiver than NCLB. Again, these gains for low-performing math students during the waiver are similar in magnitude to the differential discontinuities estimated in Chapter 2.

The negative effects for high-performing math students (β= -0.021, $p<0.001$) during NCLB were not found in the previous chapter. In contrast to the previous chapter, the definition

[13] This analysis differs from Chapter 2 in two ways: (a) the labels used (Below Basic/Basic/Proficient/Advanced compared to MYP/Priority/Enrichment) and (b) the model (local linear RD compared to DID). The results from Chapter 2 were tested for sensitivity by running the DID model from Equation 4 on the data and labels from Chapter 2 (i.e., substituting the binary variables of Below Basic, Proficient, and Advanced for the new district-provided labels of MYP and Enrichment). The previous chapter's results are generally confirmed when using the DID methods (shown in Appendix Table B1) instead of local linear RD. Students labeled Below Basic on the later benchmarks score significantly worse than Basic students during NCLB, with those losses mitigated after the waiver was implemented. The estimates are similar in magnitude across the two methods. While the significant results from Chapter 2 were affirmed here, using the DID models resulted in additional significant differences between performance labels. These labels were relatively small in magnitude, however, and similar in size to the estimated effects found in Ch. 2. For example, in reading, the model in Equation 4 detected a significant negative effect of the Proficient label on Benchmark B (β= -0.017, $p<0.05$). The estimated discontinuities for students close to this threshold from Chapter 2 were very similar (ranging from -0.015 to -0.021). The increased significance is likely due to the smaller standard errors resulting from the larger sample used in the DID models, which includes all students rather than limiting the sample to students around the threshold.

of high performers used here groups students near the Advanced threshold together. This provides evidence that both low- and high-performers did worse during NCLB than bubble students (i.e., that schools engaged in triage). The waiver does not appear to have benefitted high performers. Students labeled Enrichment on Benchmarks A and B gained significantly less during the waiver than NCLB. Overall, using a different definition for bubble students does not alter the conclusion from Chapter 2 that schools practiced less triage during the waiver.

The reading results from this analysis differ more from those in Chapter 2 than the math results. The previous chapter found some evidence that Below Basic students close to the Basic threshold on the third reading benchmark gained less during NCLB. The current results indicate that during NCLB, low performers on the third reading benchmark gained significantly more than bubble students (β=0.033, $p<0.001$). These results, then, do not indicate that schools focused on this group of bubble students in reading during NCLB. Given that this model found low-achievers in reading gained more than bubble students during NCLB, it is not particularly surprising that no differential effect is detected for being labeled MYP in reading during the waiver. Chapter 2 found that low-performing reading students benefitted during the waiver, but those gains occurred in models which showed that those students had been harmed during NCLB. There are no differences in outcomes detected in either era for high-performing students in reading.

Discussion

This analysis examined a second alternative for the reduced focus on bubble students found in Chapter 2: schools continued to triage under the waiver, but they defined bubble students differently. Rather than using the Below Basic, Basic, Proficient, and Advanced test-score labels from Chapter 2, this analysis substituted the district-supplied prescriptive labels of

MYP, Priority, and Enrichment. The math results are similar to those found in the previous chapter: (a) schools engaged in triage in math during NCLB and (b) low-performers in math benefitted from the waiver. These results further strengthen the case that the reduced focus on bubble students found in Chapter 2 was real and not an artifact of the way bubble students were defined.

So far, analyses have compared outcomes for students (a) across different incentives (both school accountability and educator evaluation systems), (b) using different definitions of bubble students, and (c) running different statistical models. The results thus far indicate that low-performers benefit under new incentives. Both the new educator evaluation system and the waiver accountability system changed the incentives facing educators, and low-achievers in GCPS gained after these changes. The gains for low-performers after these policy shifts, however, counteract previous harm done to them. These results demonstrate that policy incentives matter. That the negative effects found so far are focused on low-performers should be concerning for policymakers who intend accountability policy to support the learning of all students.

Alternative 3: Additional Funding Provided by the District Mitigated the Need to Divert Resources

The first two analyses in this chapter suggest that the reduced focus on bubble students found in Chapter 2 was due to, at least in part, shifts in the incentives of the educator evaluation system but not due to changes in how schools defined bubble students. The third hypothesis relates to the additional TAIP funding the district provided to 29 schools in 2012-13. If these schools had been engaging in triage by diverting resources towards bubble students, then the additional resources provided by the TAIP could limit triage because bubble students could be targeted without having to divert resources from other students.

The TAIP funding is investigated in two ways. First, in this chapter, I modify the previous chapter's models to control for whether a school would receive TAIP funding in 2012-13. Additional variables and interactions are included to allow the TAIP money to moderate the effect of treatment (i.e., receiving the higher of two labels). This investigates whether the TAIP funding helps explain the gains found for low-performing students during the waiver. Second, the TAIP documents are analyzed in-depth in the next chapter. That analysis reveals which students were targeted with what resources, information that gets inside the black box and answers directly whether TAIP funds were used to support low-performing students.

Data

The student-level data from the previous chapter is used. The 29 schools that received additional funds from the district filled out TAIPs, which describe their interventions for restructuring the school day. While these proposals are analyzed in-depth in the next chapter, this analysis investigates whether students who attended schools that would receive additional funds in 2012-13 had systematically different outcomes from students who attended schools that would not receive funds. To that end, a binary variable (*HasTAIP)* equals one for students who attended one of the 29 TAIP-funded schools. This variable takes a value of one across all years in the sample for students who attended these schools. In addition, because 2012-13 was the only year that TAIP funding was provided to schools, an indicator variable *2013* is included in several interactions.

Because the Chapter 2 results were concentrated at the Basic and Proficient thresholds on Benchmarks B and C, these analyses are limited to those thresholds and benchmarks.

Methods

To test the hypothesis that the increased focus on low-performers during the waiver was due to the TAIP funding given to 29 schools in 2012-13, the previous chapter's models are modified to interact the treatment variables with variables indicating (a) whether students attended schools that would receive TAIP funding in 2012-13 and (b) whether the year is 2012-13. This further modifies the RD model so that it is a difference-in-difference-in-differences (DIDID) comparison across TAIP schools, waiver, and year. Even though TAIP funding was offered to schools only in 2012-13, the entire sample—from 2009-10 through 2013-14—is included in this model. The 2013-14 school year was during the waiver, but schools did not receive TAIP funding in that year. Including both a waiver (which has a value of one for both 2012-13 and 2013-14) and a 2013 indicator in this model allows for estimating whether there continues to be a waiver effect after controlling for the TAIP funding.

Like the previous chapter, the sample is limited on each benchmark to students who received the labels on either side of a threshold, and then local linear RD models are run to estimate the discontinuities in outcomes for students who were assigned the higher of the two labels across varying bandwidths. The previous chapter included a treatment modifier (*T x waiver*) to estimate the differential effect of receiving the higher label during the waiver; this analysis includes the two additional treatment modifiers as shown below in Equation 5.

$$\begin{aligned} Y_{sijt} = {} & \beta_1 T_{sijt} + \beta_2 (T_{sijt}\ x\ HasTAIP_j) + \beta_3 (T_{sijt}\ x\ Waiver_t) + \beta_4 (HasTAIP_j\ x\ Waiver_t) + \\ & \beta_5 (T_{sijt}\ x\ HasTAIP_j\ x\ Waiver_t) + \beta_6 (T_{sijt}\ x\ 2013_t) + \beta_7 (HasTAIP_j\ x\ 2013_t) + \\ & \beta_8 (T_{sijt}\ x\ HasTAIP_j\ x\ 2013_t) + \beta_9\ num\_correct_{sijt} + \beta_{10} (T_{sijt}\ x\ num\_correct_{sijt}) + \\ & \delta X_{it} + \gamma Y_{jt} + \eta_j + \theta_t + e_{ijt} \end{aligned} \tag{5}$$

In this model, Y_{sijt} represents the standardized end-of-year state test score on the subject-area test *s* for student *i* in school *j* in year *t*. T_{sijt} represents treatment (a binary variable indicating that the student received the higher of the two performance labels in subject *s*), $HasTAIP_t$ is a binary variable indicating that the school would receive funding in 2012-13, $Waiver_t$ is a binary variable indicating that the state was under the waiver accountability system in year *t*, 2013_t is a binary variable indicating that the year was 2012-13, $num\_correct_{sijt}$ represents the distance from the threshold (students who score below the cut-score have negative values, students who score at or above have zero or positive values), X_{it} is a vector of student controls (for student *i* in year *t*, including prior year state test score, race, FRPL, ELL, and disability status), Y_{jt} is a vector of time-varying school controls (for school *j* in year *t*, including log enrollment, prior year's percentage of students scoring proficient in subject *s*, and the percentage of students who are Black, Hispanic, White, FRPL, ELL, and have a disability), η_j are school fixed effects, θ_t are grade-by-year fixed effects, and e_{ijt} is the error term. Standard errors are clustered at the school by grade level.

The omitted group in this model are the students who received the lower of the two labels during NCLB who attended schools that would not receive TAIP funding. The β_1 and β_3 are the same estimates from Chapter 2, which represent the effect of receiving the higher label during NCLB (*T*) and the differential effect during the waiver (*T x waiver*), respectively. Because the TAIP funding and year are controlled for in this analysis, the β_3 estimates (*T x waiver*) are of main interest in this model to test the hypothesis that the TAIP funding explains the previous chapter's results. If the β_3 estimates have the same direction and significance as those in the previous chapter, then the TAIP funding does not explain the shift in focus to low-performers. If the β_3 estimates are of a different sign or zero, then that suggests that the TAIP funding contributed to the gains for low-achievers.

While the β_3 coefficient relates to the hypothesis being tested in this chapter, Equation 5 also allows for investigating several other relationships of interest regarding the TAIP funding, especially to preview analyzing the actual TAIPs in the next chapter. As will be explained in more detail in Chapter 4, the district offered TAIP funding to schools and encouraged them to target students labeled as Priority with academic interventions (i.e., the students on the bubble around proficiency). Even though the TAIP funding was offered in the first year of the waiver and the second year of the new evaluation system—policy changes that my earlier analyses indicated benefitted Below Basic students—a number of TAIP schools offered these interventions to Basic students[14]. In the context of this analysis, then, the threshold where students were most likely targeted was at the Basic threshold, with Basic students receiving TAIP interventions, and Below Basic students being left out of them. This threshold is already of interest due to the main effects from the previous chapter being concentrated here. At this threshold, the Basic students are the treated students and have a value of one for the treatment (*T*) variable.

Did targeted students benefit from the TAIP funding? If schools targeted Basic students with the TAIP interventions, and if Basic students benefitted from those interventions, then Basic students in 2012-13 in TAIP schools should have better outcomes than similar students in those schools the following year, when no additional funding was available. The most straightforward way to answer this question is to look at the coefficient on *T x HasTAIP x 2013* (β_8). This DIDID

[14] Evidence for this will be forthcoming in Chapter 4.

estimator represents the difference in average outcome for Basic students in TAIP schools in 2012-13 and 2013-14 after accounting for the differences between (a) Basic students in those two years in non-TAIP schools (*T x 2013*) and (b) Below Basic students in those two years in TAIP schools (*HasTAIP x 2013*). If the β_8 estimates are positive, then Basic students had better outcomes than similar students in the same schools when they did not have additional funding. If the difference is zero, then students in both groups had similar average outcomes, meaning the TAIP funds did not benefit the Basic students. This DIDID estimator also represents the difference between treated Basic students and non-treated Below Basic students in TAIP schools during 2012-13, which is another way of indicating whether the TAIP targeting affected outcomes.

Did non-targeted students benefit from the TAIP funding? If the TAIP funding allowed schools to target the Basic students without having to take resources from Below Basic students, then non-targeted students in TAIP schools in 2012-13 (when TAIP funding was available) should do better than non-targeted students in those schools the following year (when there were no additional funds). This question is answered by the β_7 coefficient (*HasTAIP x 2013*), which represents the difference in average outcome between Below Basic students in the TAIP year with similar students the following year. A positive value for β_7 would provide evidence that the TAIP money allowed schools to target Basic students without harming Below Basic students.

The Basic threshold is of more interest in this analysis than the Proficient threshold because both (a) there was less evidence in Chapter 2 of a difference in treatment at the Proficient threshold in either subject, and (b) the second alternative explored in this chapter indicates that this might be because schools treated students around this threshold similarly (i.e., viewed them as bubble students). Because of these reasons, the main differences in outcomes are expected at the Basic threshold, although the Proficient threshold is still explored.

At the Proficient threshold, the treatment indicator refers to Proficient students. At this threshold, the students labeled as Basic have a zero value for *T,* although that does not mean they are non-targeted (Basic students just below proficiency are likely targeted by schools), it simply means they received the lower of the two labels. Because of this, β_8 (*T x HasTAIP x 2013*) refers to the DIDID estimate between Proficient students in TAIP schools in 2013 and both (a) Proficient students in TAIP schools in 2014 and (b) Basic students in TAIP schools in 2013. These students may have been targeted or schools may have viewed them as “safe” cases, making the comparison between them (i.e., β_8) less interesting than the comparison between students just below that threshold (i.e., β_7, *HasTAIP x 2013*). The β_7 value represents the difference in average outcome between Basic students in TAIP schools in the funded and following years. This coefficient would be positive if the TAIP-funding was targeted at Basic students and was effective.

Results

Math. Table 10 shows the results for students who scored just above and below the Basic threshold on the second and third math benchmarks. The estimated discontinuity at this threshold on Benchmark C during NCLB is significant and positive (ranging from 0.024 to 0.039, $p<0.05$). These estimates are similar in magnitude to those from the previous chapter and affirm that schools focused more on math students labeled Basic than those labeled Below Basic during NCLB.

The previous chapter found that Below Basic students gained during the waiver, with significant negative coefficients at this threshold for the *T x waiver* interaction on both

Benchmark B (β≈ -0.03) and Benchmark C (β≈ -0.06). Those estimates are compared with the same interaction in the new model (i.e., β_3, *Basic x waiver*) to see if the TAIP funding explains those results. The *Basic x waiver* estimates in Table 10 are consistently negative but somewhat smaller in magnitude than the previous chapter's estimates, with Benchmark B ranging from -0.019 to -0.028, and Benchmark C ranging from -0.031 to -0.050. The standard errors in these models are much larger than those from Chapter 2, meaning these estimates are not statistically significant. Because the magnitude and direction of these estimates are similar to the previous results, however, this analysis indicates that the TAIP funding does not explain the gains for Below Basic students during the waiver.

How did the additional resources affect students just above the Basic threshold during 2013, when the funding was available, compared to 2014, when it was not? The DIDID estimates of *Basic x HasTAIP x 2013* on Benchmark C are relatively large and positive (ranging from 0.050 to 0.068), indicating that Basic students in TAIP schools in 2013 did better compared to (a) Basic students in TAIP schools in 2014 and (b) Below Basic students in TAIP schools in 2013. This suggests that Basic students benefitted from the TAIP funds, although these estimates are imprecisely measured on this benchmark and inconsistent for Benchmark B.

Given that the Basic students did better in 2012-13 (suggesting that the TAIP-targeting benefitted those students), did the non-treated (Below Basic) students in TAIP schools do better in 2012-13 than similar students in 2013-14? The *HasTAIP x 2013* estimates in Table 10 are consistently negative across benchmarks, with the majority of them ranging from -0.042 to -0.102. Although these estimates are imprecisely measured, they are in the opposite direction than was hypothesized; the negative coefficients indicate that these lower-performing students did worse in the year that funding was available than they did the following year. This is counterintuitive because the additional funds should have allowed schools to provide Basic students with special treatment without harming Below Basic students and leads to questions about how TAIP funds were spent.

Table 11 shows the estimates for students close to the Proficient threshold on Benchmarks B and C. The *T x waiver* coefficients at this threshold in Chapter 2 were less stable than at the Basic threshold. The estimates were consistently negative across bandwidths (between -0.02 and -0.03) but statistically significant only at the wider bandwidths. These results suggested that schools shifted their focus during the waiver from Proficient to Basic students. The estimates in Table 11 for *Proficient x waiver* are similar in sign and magnitude on Benchmark C, meaning the TAIP funding does not account for the gains for Basic students (over Proficient students).

How did the TAIP-funding affect the Proficient students (i.e., the students who have a value of one for *T* in this table)? The estimated difference between Proficient students' average outcomes in 2013 and 2014 in TAIP schools (*Proficient x hasTAIP x 2013)* are consistently positive across both benchmarks (ranging from 0.040 to 0.099). The estimates are larger in magnitude and significant on the Benchmark B, meaning that students who scored just above proficiency did better when their school had TAIP funds than the following year. This triple difference estimator also represents the DIDID estimator between Proficient and Basic students in the TAIP-funded year. That the Proficient students scored significantly higher than Basic students is somewhat puzzling, given that the Basic students were likely targeted as well.

The comparisons between the students labeled Basic who were just below proficiency in TAIP schools in 2013 and 2014 are labeled *HasTAIP x 2013* in Table 11. These estimates are consistently negative across benchmarks, indicating that Basic students just below proficiency did worse when their schools had TAIP funding than the following year. Once again, this is a counterintuitive finding and leads to questions about how schools spent the TAIP funds.

Reading. The reading effects found in Chapter 2 were more narrowly focused than those in math. There was only some evidence that schools focused more on Basic students than Below Basic students during NCLB, with consistently positive estimates on Benchmark C (although those estimates were significant only on the higher bandwidths). This is also the only threshold and benchmark that showed a consistent shift in focus to Below Basic students during the waiver ($\beta \approx -0.045$). Table 12 includes the results from Equation 5 at the Basic threshold for reading Benchmarks B and C. Although they are not significant, the estimates for β_3 (*T x waiver)* are similar in magnitude to those from the previous chapter, with most of them between -0.04 and -0.05. This indicates that the TAIP funding does not explain the gains for students labeled Below Basic on the third reading benchmark.

The effects of the TAIP funding on reading students are now considered. The DIDID estimates (*Basic x HasTAIP x 2013*) are consistently positive across both benchmarks and larger in magnitude on Benchmark C (ranging from 0.090 to 0.136) than on Benchmark B (ranging from 0.015 to 0.055). These positive estimates indicate that students who scored just above the Basic threshold did better when their school got TAIP funds than the following year. They also indicate that the TAIP-targeted students did better than the non-targeted students during the funded year. These positive estimates were also found in math and suggest that schools used TAIP funds to target Basic students, with a positive effect on state test scores.

The final comparison is in TAIP schools between Below Basic students in 2012-13 and in 2013-14 (i.e., the non-treated students, represented by *HasTAIP x 2013* in Table 12). While these students should have better outcomes in 2012-13 if the TAIP resources allowed schools to target

Basic students without diverting resources from Below Basic students, the estimates are consistently negative across both benchmarks. The Benchmark C results are relatively large in magnitude and statistically significant across all bandwidths (ranging from -0.132 to -0.165, $p<0.05$). These estimates indicate that the students barely labeled Below Basic did significantly worse in in 2012-13, when their schools had additional TAIP funds, than in 2013-14. A similar result was found in math which suggests that schools used the TAIP funding in ways that harmed low-performers[15].

Discussion

The main purpose of this analysis was to explore the hypothesis that the shift in focus to lower-performing students found in Chapter 2 was due to the TAIP funds. The results do not support this hypothesis. Even when controlling for the TAIP funds and year, there is still evidence that Below Basic students in both math and reading benefitted during the waiver. Neither do the TAIP funds explain the leftward shift found on math Benchmark C at the Proficiency threshold.

The effects of the TAIP funding on outcomes went against expectation for the non-targeted students. In both math and reading, students who were barely labeled Below Basic had worse outcomes in the year their schools received funding than the following year. These students were hypothesized to do better in 2012-13 because the additional TAIP resources would allow schools to target the Basic students without having to divert resources from other (Below

[15] The reading results for the Proficient threshold are shown in Appendix Table B3 because Chapter 2 found no consistent significant differences in average outcome at this threshold. Appendix Table B3 shows that students who score just above proficiency had better state test scores during the TAIP-funding year than the following year (i.e., *Proficient x HasTAIP x waiver* estimates are positive). This suggests that these students were included in the targeted interventions and that they were effective. Conversely, however, the students who score just below proficiency on reading Benchmark B have worse outcomes in the TAIP year than in 2013-14 (i.e., *HasTAIP x 2013*). Because these students were likely targeted with interventions, it is again confusing that these students would have worse outcomes when the TAIP funds were available.

Basic) students. The positive coefficients on the DIDID estimator across subjects indicates that schools did in fact target the Basic students and that the targeting had a positive effect on their state test scores. Yet the targeting appears to have harmed low-performers. These results raise questions about how schools were using TAIP funds. Luckily, the TAIP documents analyzed next include precisely that information.

Conclusion

Chapter 2 reported that schools focused more on bubble students during NCLB and shifted attention towards lower-achieving students after the waiver was implemented. These results were stronger in math than in reading, and strategic behavior occurred after the third benchmark results. This chapter examined several alternate hypotheses for these results.

The first alternative hypothesized that it was not the change in incentives from the waiver accountability system which shifted behavior but instead the inclusion of value-add scores to evaluate educators. This was investigated by exploiting the fact that the new evaluation policy was implemented a year before the waiver took effect. This analysis found significant gains for Below Basic students beginning in 2011-12, which indicates that the new evaluation system was a contributing factor to the gains for lower-performing students.

The second hypothesis posited that schools used a different definition of bubble students than the test-score labels analyzed in Chapter 2. This hypothesis was examined by using the district-supplied Priority label to identify bubble students, as well as those who were too low or too high. The results are similar to those from Chapter 2: low-achieving students gain less during NCLB and rebound during the waiver. The evidence is stronger in math than in reading.

These results consistently indicate that low-performers benefit under new incentives, but these increases occur because they counteract harm done to these students during NCLB. These results indicate that (a) incentives matter and (b) the negative effects are focused on low performers.

The final alternative hypothesized that additional funding in 2012-13 allowed schools to focus on bubble (Basic) students without needing to divert resources from low-performers (Below Basic). The improved performance for low-performers was found even when the Chapter 2 models were modified to control for the TAIP funds and year.

A limitation of these black box analyses is that the mechanisms which contribute to the differences in relative test gains for students across these groups cannot be identified. Without knowing what schools did to target students, it is difficult to interpret the results. In a handful of schools, however, there is an opportunity to peek inside the black box. The TAIP proposals investigated in Chapter 4 provide insight into principals' views of how best to spend the TAIP funds, including which students schools planned to target and what resources would be provided to them.

Tables

Table 7. Estimated discontinuities at label thresholds before and after adoption of a new educator evaluation system, math

		Below Basic/Basic					Basic/Proficient					Proficient/Advanced				
	Bandwidth (number of questions) →	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
Benchmark A	T	-0.002	-0.000	-0.018	-0.017	-0.007	-0.004	0.019	0.007	0.013	0.020	-0.010	-0.038	-0.014	-0.008	-0.004
		(0.024)	(0.017)	(0.014)	(0.013)	(0.013)	(0.022)	(0.016)	(0.014)	(0.013)	(0.012)	(0.030)	(0.023)	(0.020)	(0.018)	(0.017)
	T x Eval	0.008	0.006	0.015	0.020	0.015	-0.036*	-0.031	-0.030	-0.028	-0.027	0.019	0.008	-0.011	-0.022	-0.018
		(0.019)	(0.017)	(0.017)	(0.016)	(0.016)	(0.018)	(0.016)	(0.016)	(0.015)	(0.015)	(0.025)	(0.021)	(0.019)	(0.019)	(0.019)
	Obs	20247	27160	31869	34701	36196	17072	23730	29798	33493	35445	7555	10690	14128	17385	18977
	R^2	0.447	0.469	0.490	0.506	0.516	0.492	0.522	0.554	0.567	0.573	0.420	0.466	0.510	0.542	0.560
Benchmark B	T	-0.038	-0.007	0.005	0.019	0.021	-0.006	0.002	0.007	0.016	0.017	0.086**	0.028	0.028	0.000	-0.007
		(0.023)	(0.017)	(0.014)	(0.013)	(0.012)	(0.021)	(0.016)	(0.013)	(0.012)	(0.012)	(0.028)	(0.021)	(0.019)	(0.017)	(0.016)
	T x Eval	0.016	0.012	0.003	-0.001	0.004	-0.026	-0.020	-0.015	-0.004	-0.002	-0.039	-0.032	-0.019	-0.008	-0.004
		(0.022)	(0.018)	(0.017)	(0.016)	(0.016)	(0.017)	(0.015)	(0.014)	(0.014)	(0.014)	(0.024)	(0.022)	(0.020)	(0.020)	(0.019)
	Obs	17060	23184	28193	31415	33389	17189	23658	29542	33619	35498	9017	12760	16477	19714	21443
	R^2	0.421	0.447	0.476	0.501	0.516	0.485	0.518	0.548	0.563	0.567	0.460	0.499	0.529	0.563	0.582
Benchmark C	T	0.006	0.024	0.038**	0.042**	0.040**	-0.007	-0.012	-0.019	-0.011	-0.003	0.054	0.028	0.030	0.027	0.022
		(0.022)	(0.017)	(0.014)	(0.013)	(0.012)	(0.019)	(0.014)	(0.013)	(0.011)	(0.011)	(0.028)	(0.021)	(0.018)	(0.016)	(0.016)
	T x Eval	-0.010	-0.036*	-0.043*	-0.045**	-0.041**	-0.011	-0.007	-0.006	-0.008	-0.006	0.017	0.026	0.023	0.025	0.025
		(0.021)	(0.018)	(0.017)	(0.016)	(0.016)	(0.016)	(0.014)	(0.014)	(0.013)	(0.013)	(0.021)	(0.019)	(0.017)	(0.017)	(0.017)
	Obs	16509	22461	27968	31728	33261	15841	21834	27690	32285	35009	9089	12756	16401	19859	21561
	R^2	0.412	0.447	0.486	0.508	0.517	0.437	0.478	0.528	0.559	0.576	0.458	0.499	0.537	0.571	0.592

Notes: *p<0.05; **p<0.01, ***p<0.001. "T" represents the discontinuity in outcomes for students who received the higher of the two labels. "T x Eval" represents the difference in the discontinuity during 2011-12, when the new educator evaluation system was implemented. Standard errors in parentheses are clustered at the school by grade by year level. All models condition on student covariates (race, ethnicity, FRPL, ELL, and disability status), school covariates (log enrollment, prior year's percent proficient in math, percent of school that is black, Hispanic, white, FRPL, ELL, and has a disability), grade by year fixed effects, and school fixed effects.

Table 8. Estimated discontinuities at label thresholds before and after adoption of a new educator evaluation system, reading

		Below Basic/Basic					Basic/Proficient					Proficient/Advanced				
	Bandwidth (number of questions) →	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
Benchmark A	T	0.011	-0.000	0.016	0.012	0.018	0.004	0.001	0.016	0.021	0.021*	0.040	0.024	0.016	0.026	0.022
		(0.032)	(0.022)	(0.019)	(0.017)	(0.016)	(0.022)	(0.015)	(0.013)	(0.011)	(0.010)	(0.025)	(0.019)	(0.016)	(0.015)	(0.014)
	T x Eval	0.020	-0.008	-0.013	-0.021	-0.031	0.009	0.002	-0.005	-0.009	-0.010	-0.012	-0.022	-0.020	-0.017	-0.013
		(0.025)	(0.021)	(0.018)	(0.017)	(0.017)	(0.017)	(0.015)	(0.014)	(0.013)	(0.012)	(0.023)	(0.019)	(0.018)	(0.017)	(0.017)
	Obs	10111	14003	17725	21259	24615	15137	20861	26088	31265	35200	10008	13795	17174	20257	23136
	R^2	0.405	0.425	0.440	0.464	0.487	0.450	0.485	0.522	0.558	0.585	0.438	0.480	0.519	0.546	0.568
Benchmark B	T	0.025	0.006	0.007	-0.001	0.017	-0.033	-0.034*	-0.025	-0.018	-0.018	-0.003	-0.004	0.015	0.018	0.020
		(0.028)	(0.020)	(0.018)	(0.016)	(0.015)	(0.021)	(0.015)	(0.013)	(0.011)	(0.010)	(0.025)	(0.018)	(0.016)	(0.014)	(0.013)
	T x Eval	-0.000	-0.008	-0.024	-0.037*	-0.043**	-0.011	0.001	0.000	-0.012	-0.013	0.014	0.002	-0.006	-0.009	-0.017
		(0.022)	(0.020)	(0.018)	(0.017)	(0.017)	(0.016)	(0.014)	(0.013)	(0.012)	(0.011)	(0.022)	(0.019)	(0.018)	(0.018)	(0.018)
	Obs	10382	14399	18225	21522	24258	15493	21414	27029	32131	35781	10643	14643	18182	21552	24524
	R^2	0.372	0.406	0.436	0.465	0.482	0.435	0.478	0.521	0.561	0.592	0.456	0.491	0.520	0.551	0.572
Benchmark C	T	0.043	0.032	0.031	0.042**	0.040**	0.008	-0.004	0.003	0.000	-0.002	0.034	0.021	0.019	0.009	0.016
		(0.030)	(0.020)	(0.018)	(0.016)	(0.015)	(0.020)	(0.015)	(0.012)	(0.011)	(0.010)	(0.025)	(0.018)	(0.016)	(0.015)	(0.014)
	T x Eval	-0.009	-0.033	-0.051**	-0.060***	-0.065***	-0.022	-0.026	-0.031*	-0.025*	-0.019	-0.035	-0.028	-0.021	-0.017	-0.019
		(0.025)	(0.020)	(0.019)	(0.018)	(0.018)	(0.016)	(0.014)	(0.012)	(0.012)	(0.011)	(0.020)	(0.020)	(0.019)	(0.019)	(0.019)
	Obs	9851	13878	17711	21565	24770	15914	21902	27724	32761	36744	9832	13412	16696	19865	22483
	R^2	0.429	0.449	0.474	0.500	0.515	0.443	0.485	0.534	0.575	0.607	0.450	0.489	0.523	0.550	0.565

Notes: *p<0.05; **p<0.01, ***p<0.001. “T” represents the discontinuity in outcomes for students who received the higher of the two labels. “T x Eval” represents the difference in the discontinuity during 2011-12, when the new educator evaluation system was implemented. Standard errors in parentheses are clustered at the school by grade by year level. All models condition on student covariates (race, ethnicity, FRPL, ELL, and disability status), school covariates (log enrollment, prior year’s percent proficient in reading, percent of school that is black, Hispanic, white, FRPL, ELL, and has a disability), grade by year fixed effects, and school fixed effects

Table 9. Regression results by prescriptive label during NCLB and waiver

	Math			Reading		
	Benchmark A	Benchmark B	Benchmark C	Benchmark A	Benchmark B	Benchmark C
Multi-Year Plan	-0.007	-0.019*	-0.030***	-0.005	-0.011	0.033***
	(0.010)	(0.010)	(0.009)	(0.009)	(0.009)	(0.009)
Enrichment	0.003	0.001	-0.021*	0.015	0.018	-0.003
	(0.013)	(0.011)	(0.011)	(0.010)	(0.010)	(0.010)
Multi-Year Plan x Waiver	0.002	0.006	0.043***	-0.005	-0.011	-0.011
	(0.012)	(0.013)	(0.011)	(0.009)	(0.009)	(0.009)
Enrichment x Waiver	-0.062***	-0.031*	0.006	0.016	-0.010	0.002
	(0.015)	(0.014)	(0.013)	(0.012)	(0.012)	(0.011)
Observations	122185	125035	126349	126172	128230	127436
Adjusted R^2	0.64	0.67	0.71	0.71	0.73	0.74

Notes: *p<0.05; **p<0.01, ***p<0.001. NCLB includes 2009-10 through 2011-12 data. Waiver includes 2012-13 through 2013-14 data. Standard errors in parentheses are clustered at the school by grade by year level. All models condition on student testing variables up to cubic polynomials and interactions between all the variables (prior year state test z-score; projected percentile; raw score on the benchmark, centered on proficiency), student demographics (race, ethnicity, FRPL, ELL, and disability status), school covariates (log enrollment, prior year's percent proficient in that subject, percent of school that is black, Hispanic, white, FRPL, ELL, and has a disability), school fixed effects, and grade by year fixed effects.

Table 10. Estimated discontinuities at Below Basic/Basic threshold by TAIP funding and 2012-13, math

	Benchmark B					Benchmark C				
Bandwidth (number of questions) →	2	3	4	5	6	2	3	4	5	6
Basic	-0.027	-0.002	0.006	0.019	0.021	0.024	0.033*	0.038**	0.039**	0.037**
	(0.021)	(0.016)	(0.014)	(0.013)	(0.013)	(0.021)	(0.016)	(0.014)	(0.013)	(0.013)
Basic x has TAIP	0.015	0.013	0.003	0.002	0.003	0.019	0.014	0.012	0.016	0.016
	(0.018)	(0.017)	(0.016)	(0.016)	(0.016)	(0.018)	(0.016)	(0.015)	(0.014)	(0.014)
Basic x waiver	-0.023	-0.019	-0.027	-0.028	-0.028	-0.031	-0.042	-0.049	-0.050	-0.049
	(0.030)	(0.027)	(0.026)	(0.025)	(0.025)	(0.030)	(0.029)	(0.025)	(0.026)	(0.025)
Has TAIP x waiver	0.064	0.081	0.087	0.087	0.094*	0.094	0.077	0.050	0.066	0.086*
	(0.048)	(0.047)	(0.047)	(0.048)	(0.047)	(0.050)	(0.046)	(0.044)	(0.044)	(0.043)
Basic x has TAIP x waiver	-0.004	-0.021	-0.022	-0.022	-0.028	-0.027	-0.041	-0.031	-0.037	-0.052
	(0.042)	(0.035)	(0.036)	(0.035)	(0.035)	(0.045)	(0.041)	(0.037)	(0.037)	(0.036)
Basic x 2013	0.009	-0.003	0.003	0.003	0.002	-0.020	-0.022	-0.016	-0.021	-0.020
	(0.035)	(0.031)	(0.030)	(0.029)	(0.029)	(0.037)	(0.034)	(0.030)	(0.031)	(0.030)
Has TAIP x 2013	-0.014	-0.042	-0.058	-0.061	-0.074	-0.102	-0.097	-0.065	-0.078	-0.089
	(0.058)	(0.056)	(0.055)	(0.056)	(0.055)	(0.060)	(0.055)	(0.053)	(0.052)	(0.051)
Basic x has TAIP x 2013	-0.042	-0.005	0.014	0.015	0.030	0.058	0.065	0.050	0.056	0.068
	(0.050)	(0.044)	(0.043)	(0.042)	(0.041)	(0.055)	(0.049)	(0.044)	(0.044)	(0.043)
Observations	25335	34504	42531	48015	51369	22456	30638	38327	44149	47640
R^2	0.691	0.687	0.685	0.686	0.688	0.728	0.726	0.723	0.723	0.726

Notes: *p<0.05; **p<0.01, ***p<0.001. Standard errors in parentheses are clustered at the school by grade by year level. All models condition on student covariates (race, ethnicity, FRPL, ELL, and disability status), school covariates (log enrollment, prior year's percent proficient in reading, percent of school that is black, Hispanic, white, FRPL, ELL, and has a disability), grade by year fixed effects, and school fixed effects.

Table 11. Estimated discontinuities at Basic/Proficient threshold by TAIP funding and 2012-13, math

	Benchmark B					Benchmark C				
Bandwidth (number of questions) →	2	3	4	5	6	2	3	4	5	6
Proficient	-0.040*	-0.009	0.003	0.009	0.008	-0.016	-0.016	-0.020	-0.015	-0.010
	(0.018)	(0.013)	(0.012)	(0.011)	(0.011)	(0.017)	(0.013)	(0.012)	(0.011)	(0.011)
Proficient x has TAIP	0.040*	0.044**	0.037**	0.037**	0.041**	0.011	0.007	0.011	0.015	0.016
	(0.017)	(0.015)	(0.014)	(0.014)	(0.014)	(0.016)	(0.014)	(0.013)	(0.013)	(0.013)
Proficient x waiver	0.039	0.009	0.005	-0.004	-0.007	-0.009	-0.011	-0.028	-0.038	-0.037
	(0.021)	(0.020)	(0.019)	(0.018)	(0.019)	(0.026)	(0.022)	(0.020)	(0.020)	(0.020)
Has TAIP x waiver	0.111**	0.093*	0.101**	0.083*	0.078*	0.036	0.041	0.040	0.035	0.031
	(0.041)	(0.039)	(0.037)	(0.036)	(0.036)	(0.037)	(0.034)	(0.033)	(0.031)	(0.032)
Proficient x has TAIP x waiver	-0.111***	-0.088**	-0.101***	-0.077**	-0.075**	-0.059	-0.062*	-0.048	-0.037	-0.036
	(0.034)	(0.030)	(0.028)	(0.026)	(0.027)	(0.037)	(0.031)	(0.029)	(0.029)	(0.029)
Proficient x 2013	-0.044	-0.020	-0.029	-0.021	-0.013	-0.003	-0.000	0.014	0.028	0.031
	(0.026)	(0.024)	(0.023)	(0.022)	(0.023)	(0.031)	(0.025)	(0.023)	(0.023)	(0.023)
Has TAIP x 2013	-0.077	-0.051	-0.058	-0.051	-0.044	-0.022	-0.031	-0.026	-0.015	-0.012
	(0.051)	(0.047)	(0.045)	(0.043)	(0.043)	(0.048)	(0.043)	(0.039)	(0.037)	(0.037)
Proficient x has TAIP x 2013	0.099*	0.066	0.092**	0.076*	0.071*	0.053	0.069	0.062	0.042	0.040
	(0.041)	(0.038)	(0.035)	(0.033)	(0.034)	(0.046)	(0.038)	(0.035)	(0.034)	(0.034)
Observations	27476	37873	46913	53427	56835	25007	34500	43508	50579	54713
R^2	0.461	0.490	0.519	0.538	0.547	0.425	0.462	0.506	0.537	0.558

Notes: *p<0.05; **p<0.01, ***p<0.001. Standard errors in parentheses are clustered at the school by grade by year level. All models condition on student covariates (race, ethnicity, FRPL, ELL, and disability status), school covariates (log enrollment, prior year's percent proficient in reading, percent of school that is black, Hispanic, white, FRPL, ELL, and has a disability), grade by year fixed effects, and school fixed effects.

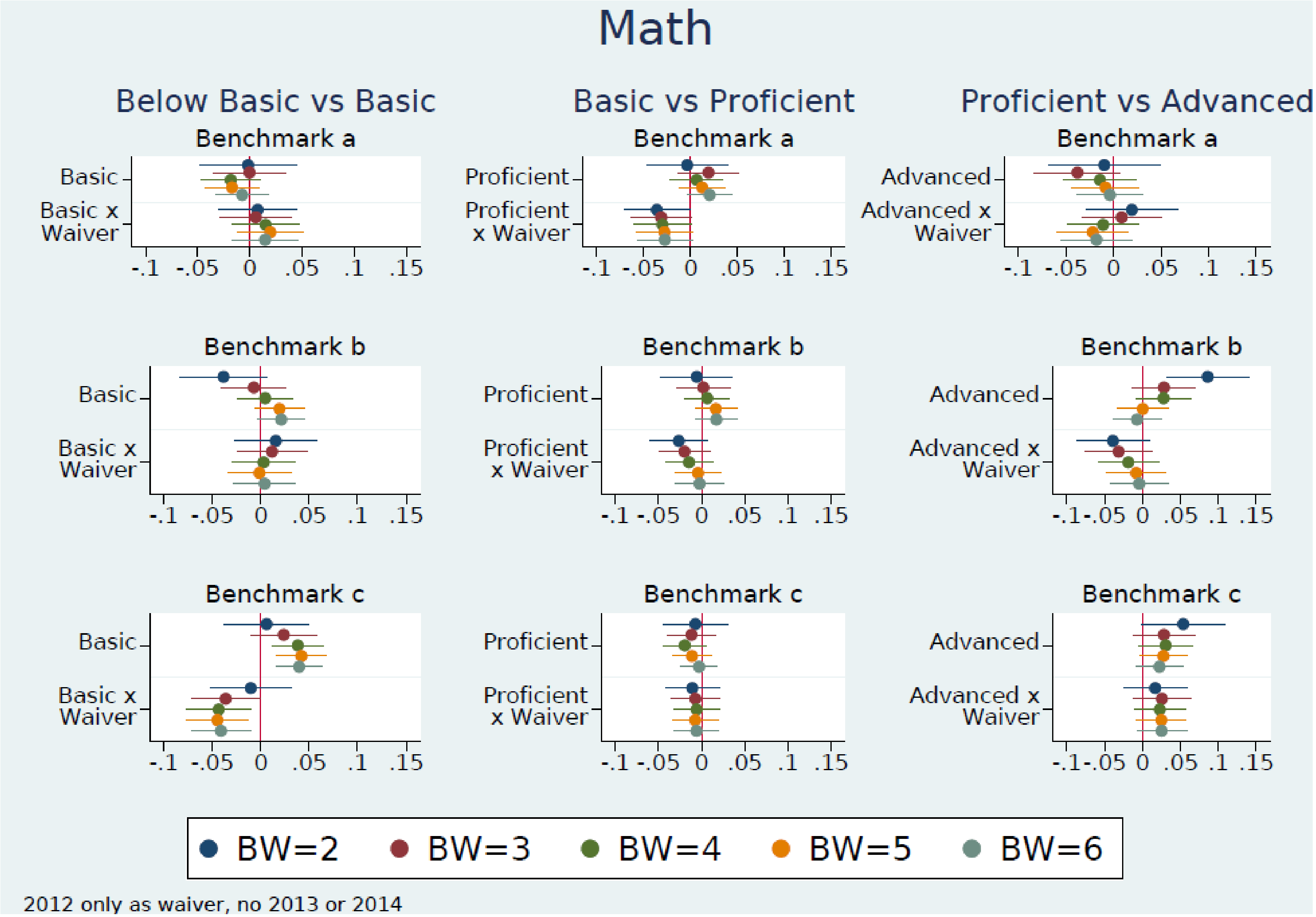
Table 12. Estimated discontinuities at Below Basic/Basic threshold by TAIP funding and 2012-13, reading

	Benchmark B					Benchmark C				
Bandwidth (number of questions) →	2	3	4	5	6	2	3	4	5	6
Basic	-0.007	-0.004	0.004	0.003	0.021	0.017	0.011	0.012	0.022	0.026
	(0.025)	(0.018)	(0.016)	(0.014)	(0.013)	(0.027)	(0.020)	(0.017)	(0.016)	(0.015)
Basic x has TAIP	0.022	0.014	0.003	0.013	0.008	0.030	0.024	0.016	0.009	0.009
	(0.021)	(0.019)	(0.018)	(0.017)	(0.017)	(0.021)	(0.018)	(0.016)	(0.016)	(0.016)
Basic x waiver	0.032	0.024	0.012	-0.012	-0.017	-0.016	-0.049	-0.043	-0.047	-0.052*
	(0.030)	(0.025)	(0.023)	(0.021)	(0.020)	(0.040)	(0.033)	(0.031)	(0.026)	(0.026)
Has TAIP x waiver	0.072	0.056	0.069*	0.055	0.056	0.090	0.089	0.087	0.082	0.086
	(0.037)	(0.034)	(0.032)	(0.032)	(0.031)	(0.066)	(0.054)	(0.051)	(0.047)	(0.046)
Basic x has TAIP x waiver	-0.039	-0.042	-0.035	-0.028	-0.031	-0.110	-0.080	-0.077	-0.065	-0.080
	(0.045)	(0.040)	(0.036)	(0.033)	(0.032)	(0.072)	(0.056)	(0.055)	(0.050)	(0.048)
Basic x 2013	-0.038	-0.036	-0.042	-0.022	-0.014	-0.022	0.028	0.004	0.004	-0.006
	(0.036)	(0.031)	(0.028)	(0.026)	(0.025)	(0.046)	(0.039)	(0.035)	(0.031)	(0.031)
Has TAIP x 2013	-0.043	-0.042	-0.076	-0.058	-0.057	-0.147*	-0.132*	-0.142*	-0.147**	-0.165**
	(0.050)	(0.043)	(0.039)	(0.038)	(0.037)	(0.074)	(0.061)	(0.057)	(0.054)	(0.053)
Basic x has TAIP x 2013	0.015	0.038	0.055	0.035	0.032	0.115	0.090	0.104	0.099	0.136*
	(0.059)	(0.049)	(0.043)	(0.040)	(0.038)	(0.081)	(0.065)	(0.061)	(0.057)	(0.055)
Observations	15675	21689	27444	32597	37199	13429	18843	24125	29452	34083
R^2	0.807	0.806	0.799	0.794	0.787	0.816	0.813	0.808	0.801	0.794

Notes: *p<0.05; **p<0.01, ***p<0.001. Standard errors in parentheses are clustered at the school by grade by year level. All models condition on student covariates (race, ethnicity, FRPL, ELL, and disability status), school covariates (log enrollment, prior year's percent proficient in reading, percent of school that is black, Hispanic, white, FRPL, ELL, and has a disability), grade by year fixed effects, and school fixed effects.

Figures

Figure 11. Estimated discontinuities at label thresholds before and after new educator evaluation system, math



2012 only as waiver, no 2013 or 2014

Figure 12. Estimated discontinuities at label thresholds before and after new educator evaluation system, reading

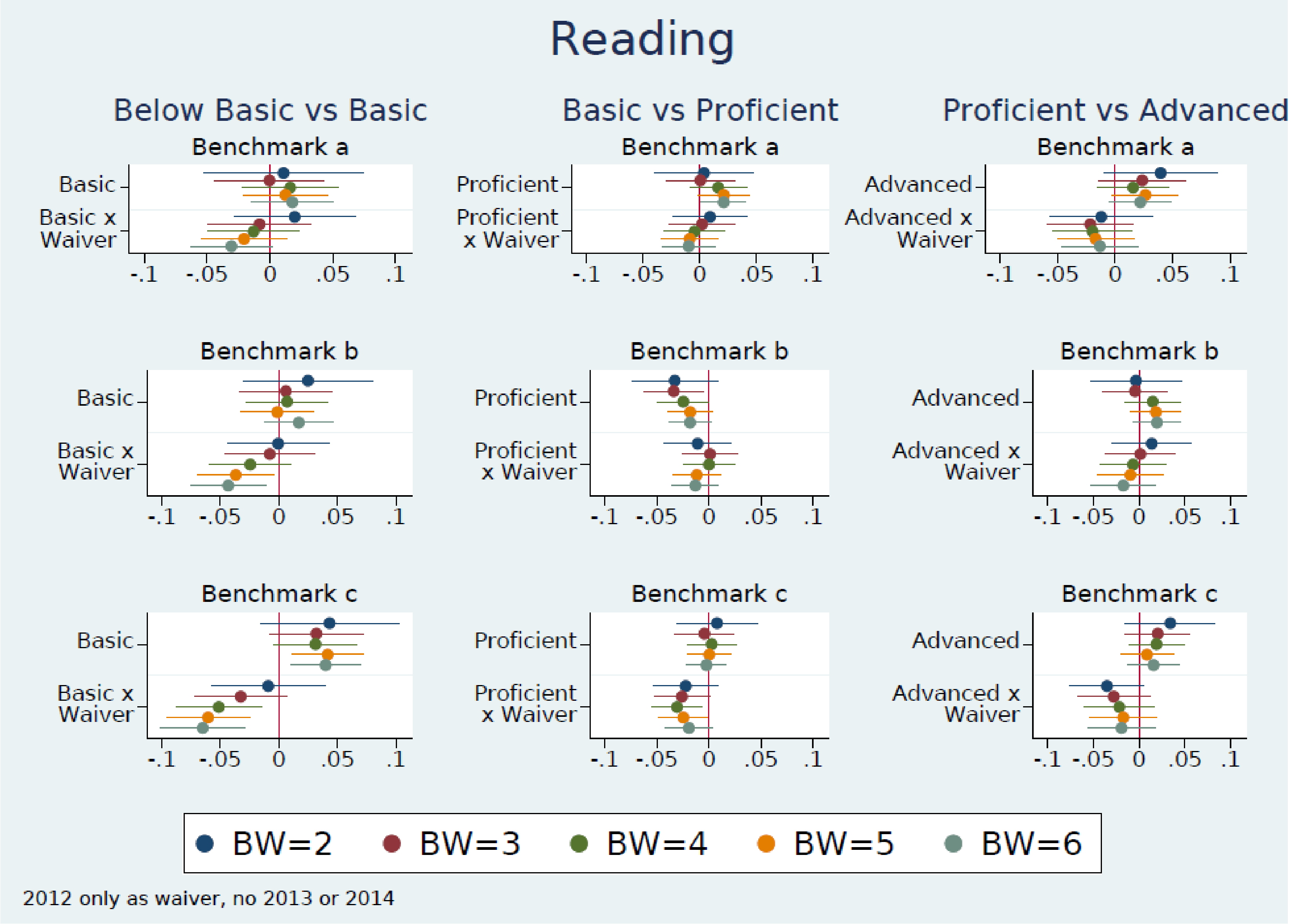


Figure 13. Percent of sample assigned each prescriptive label during NCLB and waiver, math

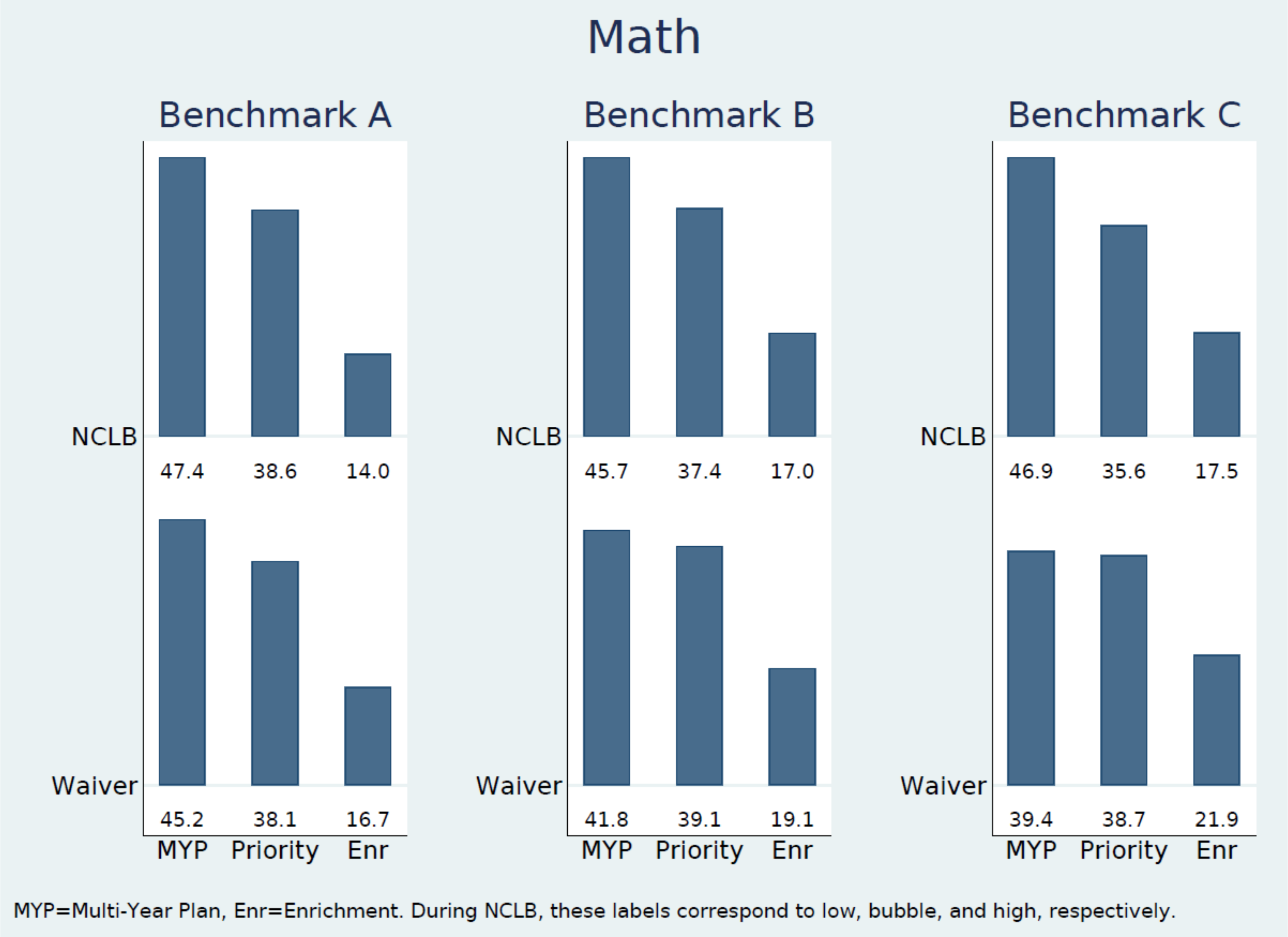
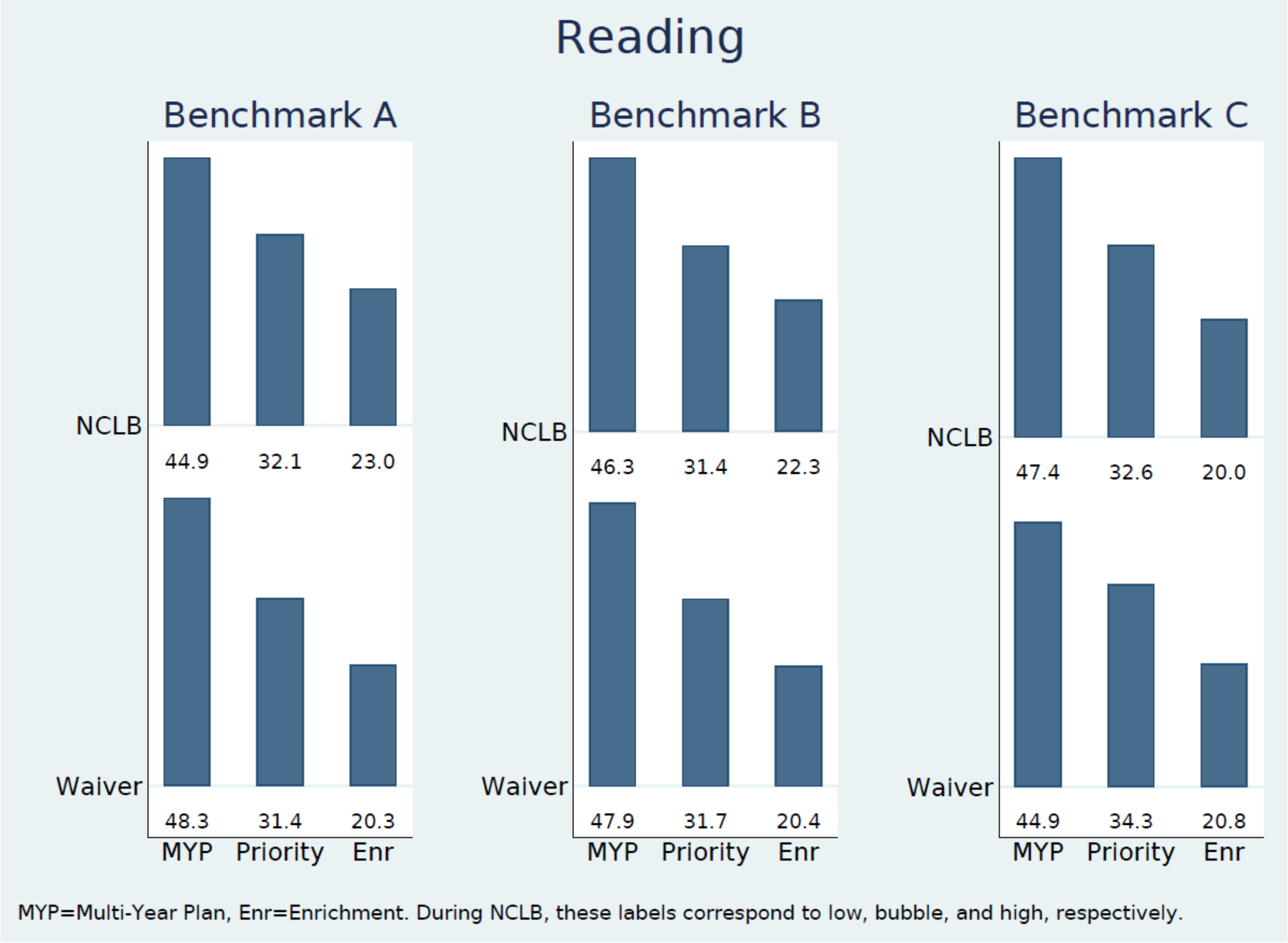


Figure 14. Percent of sample assigned each prescriptive label during NCLB and waiver, reading



Chapter 4: What Does Strategic Behavior Look Like in District-Endorsed Educational Triage?

The last alternative examined in Chapter 3 was that the additional funding for targeted academic interventions might have obviated the need for schools to engage in triage. Those results uncovered some puzzling findings regarding the 29 schools that turned in Targeted Academic Intervention Proposals (TAIPs) in 2012-13. While there was evidence that the targeted (Basic) students had better outcomes when their schools had TAIP funds, non-targeted (Below Basic) students did worse than their counterparts the following year. Those results suggest that TAIP schools used the additional funds in ways that harmed the non-targeted students.

Interpreting the previous chapter's TAIP analysis results is hampered because it is a standard quantitative black box study which does not provide information about what is actually happening in schools. The TAIP program provides an opportunity to get inside the black box because to receive the supplemental funding, school leaders in the selected schools completed TAIPs. In the TAIP documents, school leaders proposed how they would spend the money to restructure the school day to offer academic interventions for the two months before the state test. The proposals detail both the structure of the interventions and the students to whom the school planned to offer these interventions. These documents provide insight into what strategic behavior looks like in a large number of schools.

What happens when a district tells administrators which students are close to meeting proficiency and then gives them additional money to target those students? Did the schools target the Priority students or did they include more students? Did schools allocate the resources in a way that circumvented the need to triage, or did they target bubble students by diverting resources from low- and high-achievers? Given that test scores are typically used as proxies for this kind of strategic behaviors, are these school-level interventions of sufficient quality to improve test scores? This chapter uses the TAIPs in order to (a) identify the targeted students and (b) describe and analyze the features of the interventions.

RQ1: Which students did school leaders plan to target with interventions? The district pressed TAIP schools to specifically target interventions to the Priority students. The evidence for this includes the *Identification of Target Students* document shown in Figure 2. The labels, descriptors, and color-coding all align with dividing students into triage groups, with the Priority students representing the bubble students. That document specifies that schools should "prioritize *Target* students" and then labels a group of students as "Priority." In addition, the title of the TAIP document indicates that the money is intended for targeted academic interventions, which implies they are for "target" students. The consistency of this wording sends a message to school leaders that the interventions should be provided to targeted students, which are those who are labeled Priority.

Despite the district message to focus on the Priority students, school leaders could decide which students needed additional support based on their own schools' needs. Most TAIPs included information about which students were targeted and what data school leaders used to identify those students. Although the Priority students are the expected focal group, schools used several different metrics to determine which students would be targeted. The TAIPs provide information on the idiosyncratic way that each school identified their target students[16]. Exploring

[16] I want to clarify the difference between some similar terms used in this chapter to describe students (i.e., bubble, Priority, and target). As has been discussed throughout this dissertation, the term "bubble students" refers to the group of students who are close to the proficiency line, with low- and high-performers referring to students who are not "on the bubble" of proficiency. These groups are nebulously defined by researchers and by schools, and refer to broad groups of students. On the other hand, Priority students are the specific students who were assigned the

this question is important because it represents equity concerns: which students had access to additional resources and which students were left out?

RQ2: What resources were used for the targeted interventions, and how were the interventions structured? The proposed interventions represent what school leaders thought was the best way to spend additional resources in order to increase student test scores before the state test. Exploring this question provides a direct view into the black box regarding how schools reacted to student testing data and the accountability incentives in 2012-13. Knowing how principals designed these interventions offers insight into how disruptive these interventions would be at the classroom and student level.

Because quantitative researchers often lack resource allocation information at the school level, they use test scores as proxies (Springer, 2008b). The underlying assumption is that the resources allocated to students are effective in raising test scores (as shown in the logic model in Figure 1). This implicit assumption suggests two possible explanations for null quantitative results: either (a) schools are not engaging in the strategic behavior or (b) the metrics being used—test score gains—are not picking up the underlying behavior. Rather than looking at proxies, the TAIPs afford the opportunity to explore at a detailed level how school leaders planned to allocate resources.

Priority 1 or Priority 2 label from the district based on their benchmark and projected scores. These students have test scores which place them around the proficiency line, making these students the likely ones to be considered bubble by their schools (especially given the district's encouragement and financial support to focus on those students in the lead-up to the state test). A final distinction is made for target students. This term refers to the specific group of students that TAIP schools planned to target with their interventions. As just noted, each school leader determined which students would be targeted with interventions. Schools could target bubble students based on their Priority label or based on another metric, or schools could target more broadly than students close to proficiency. The term “target student” represents the most localized, school-level view of focal students in this dissertation.

To look further into these possible explanations for null results in triage work, the proposed interventions are grouped into four broad types based on features including (a) who provides the intervention, (b) when during the school day it occurs, and (c) which students are targeted (i.e., whether the interventions are equitably accessible by students in need). These dimensions are analyzed to consider whether or not the TAIP interventions might improve test scores by providing enhanced learning opportunities. Because the additional funding should allow TAIP schools to avoid needing to triage, this analysis examines whether the resources which were offered to target students came in addition to the typical resources provided to all students or whether resources were diverted from other students. This detailed analysis on a relatively large sample of schools bridges the gap between the prior triage work which uses qualitative methods on case studies of individual schools (e.g., Booher-Jennings, 2005; Horn, 2016) and that which uses quantitative methods on administrative data for large numbers of students (e.g., Ballou & Springer, 2016; Krieg, 2008; Ladd & Lauen, 2010).

Sample, Data, and Methods

Sample

Twenty-nine schools in GCPS submitted TAIPs, including 15 elementary and 14 middle schools. These TAIP schools were selected by members of the district research team who decided which schools should receive additional funds and how much they should get. The schools were selected based on the number of Priority students in the school, where the amount of funding was considered “adequate to implement a strategy” (meeting with GCPS research team). The district research team indicated that these schools were among the lowest-performing in the district.

The summary statistics shown in Table 13 compare the 29 TAIP schools with the schools in the district that did not submit a plan. Elementary schools that submitted TAIPs have slightly lower percentages of students who scored Proficient or Advanced in both subjects the previous year compared to schools without plans. The prior year performance categories are similar in the middle schools. Both middle and elementary TAIP schools have slightly higher percentages of Priority and Multi-Year Plan students than non-TAIP schools. This indicates that lower performing schools received TAIP funding.

Demographic comparisons of the schools are shown in the bottom panel of Table 13. There are fewer black students in TAIP schools than in non-TAIP schools in both elementary (30% vs 49%) and middle schools (42% vs 56%). These differences were offset by an increased percentage of Hispanic students in TAIP schools, although the percent of white students was similar across school groups. TAIP elementary schools had slightly higher percentages of students eligible for free and reduced price lunch than non-TAIP schools (83% vs 75%) as well as a substantially higher percent of English language learners (37% vs 1%). The percent of students with disabilities and the mobility rate were relatively consistent across all groups, comprising about 12% and 30% of the school, respectively. Overall, schools that received TAIP funding had lower performing students than other schools in GCPS.

Data Collection

Data for this analysis were collected through document review and student-level quantitative data. Documents included TAIPs from 29 schools, internal email communication between the district research team and school administrators, training documents from principal meetings, and examples of reports available to principals and teachers through the district data warehouse. The quantitative data were used to identify the targeted students in the TAIP schools.

Targeted Academic Intervention Proposals. Administrators at each school completed a TAIP based on the blank template shown in Figure 15. The template included a place for schools to describe how they would restructure the school day to provide targeted academic interventions. These proposals were used to create a school-level database. Two schools' TAIPs are recreated in Figure 16 to illustrate some of the methodological decisions made in these analyses.

For each intervention, the TAIP document prompted school leaders to "*Briefly describe the proposed program. (Include a description of student and teacher selection processes, a plan for communicating with stakeholders and encouraging student participation)."* Each proposal included blanks to indicate the intervention targets (grade level and subject), number of students participating, proposed beginning and end date, and how many days per week the intervention would last. The principals input their digital signature on the TAIP, and the district responded with (a) approve as requested, (b) approve with indicated modifications, or (c) not approved with additional information needed. All 29 plans were marked as "approve as requested."

Student testing data. The school-level information gathered from the TAIPs was matched with the student-level dataset used in previous chapters for students in the 2012-13 school year. This was done to triangulate the information between the TAIPs and the quantitative data in order to verify an accurate representation of which students were actually targeted[17].

[17] During this analysis, the information from the TAIPs was linked to the student-level testing data to identify which students the school proposed to actually target (i.e., treat). I attempted to quantify the test score effects of these interventions but ran into several issues. TAIP schools did not target very many students, and the standard errors on the estimates were very large. In addition, because the district does not have data on the fidelity of implementation of the interventions, I do not know the extent to which plans were implemented as written.

Data Analysis Methods

The purpose of this analysis is to explore how schools planned to use additional funding to provide interventions for students in the two months before the state test. This study is appropriate for naturalistic inquiry because it is intended to get inside the black box of quantitative research regarding school resource allocation. This work analyzes school documents, and the human instrument (Lincoln & Guba, 1985) is key in looking for patterns of school allocation decisions in those documents.

I initially coded the TAIPs along with another graduate student in education. Both of us are former teachers, and our teaching influences our perspectives and my subsequent analyses. My public education background includes experience as both a classroom math teacher and a school-based instructional coach in low-performing public middle schools during NCLB. As a math teacher, I provided interventions to struggling students that were specifically focused on increasing their state test scores. As an instructional coach, my principal tasked me with using student data to identify students for remediation, scheduling interventions during the school day, and planning the curriculum for the interventions. In that work, I completed documents—which are similar to those analyzed here—detailing my school's intervention strategies. The interventions described in the TAIPs are similar to ones I designed or knew about through conversations with instructional coaches at other schools. The other graduate student was a former English teacher, which helped provide a perspective into the other tested content area. This student had experience providing reading interventions to students in her low-performing school. Together, our education backgrounds provided insight into the quality of the interventions and whether we believed they would offer students enhanced learning opportunities.

The documents were coded utilizing the constant comparison method (Glaser & Strauss, 1967), which informed ongoing analyses based on previously collected data. While schools were expected to focus on Priority students, I took a grounded theory approach (Glaser & Strauss, 1967) to allow information learned during coding to describe alternate behaviors by school administrators. This openness to alternative intervention structures was important because (a) funded schools did not necessarily need to divert resources from one group to another (i.e., engage in triage) in order to focus on students close to proficiency, (b) schools designed interventions in a variety of ways, and (c) some schools did not use the Priority label to identify their target students.

The TAIP proposals were first coded openly (a) to identify the features of the intervention and which students were targeted and (b) to transfer those logistical features to a school-level database that could be linked with the student data. Each sentence was the unit of analysis for the open-coding. Common information taken directly from all TAIPs included the start and end dates (used to calculate the total number of school days the intervention lasted), the days per week of intervention, the grade level and subject targeted, and the number of students participating. The number of students participating was used to triangulate between the TAIP and student quantitative data in attempts to identify the actual students targeted by each school.

Open-coding revealed that the plans varied extensively in the specificity of the proposed interventions. Each school's description of their restructured school day was open-coded using logistically descriptive categories: which students were targeted, what resources were used, who provided the intervention, and when during the school day the intervention occurred. Some of the codes were generated from the language of the documents: "targeted students," "identified students," "expert teachers." These codes were initially utilized to create a spreadsheet that indicated whether the school's intervention included different features. As new features appeared, additional categories were added to the spreadsheet. If a new code emerged from a school document, the constant comparative method (Glaser & Strauss, 1967) was utilized to check the previously-coded interventions against these new codes.

The initial codes captured the language used by each school leader to describe which students were the target of the interventions and what metrics the school used to identify these students. In many cases, the schools included clear descriptors of their targeted students (e.g., School A in Figure 16, "within three answers of being proficient on the DEA assessment"). In cases where the metric the school used was ambiguous (e.g., School B in Figure 16, "targeted students"), it was assumed that Priority 1 and 2 students were targeted. That assumption was then validated by triangulating the number of Priority students found in the quantitative data with the number of students participating in the intervention that each school leader included on the TAIP description[18].

The initial coding of the TAIP documents was done individually by both researchers. After adding any new intervention features to the school database, we individually wrote an analytic memo describing the school's intervention and connections with other school proposals. These analytic memos included conjectures about the effectiveness of the intervention and explanations for why we thought the intervention would or would not increase student learning. These analytic memos allowed us to further refine which different intervention features would likely offer student enhanced learning opportunities (and, subsequently, increased test scores). In

[18] If the two values were within 8 students of each other (generally around 10% of the stated number), that school was considered a match. Figure 16 shows that School B expected 75 "targeted" students to participate in the intervention. There were 82 seventh-grade students in the quantitative data who were labeled Priority 1 or 2 in that school, making this school a match.

addition to conjectures about the quality of the plan, the analytic memos also discussed equity concerns about access to the interventions, detailing which students were being targeted and which were left out of these interventions. This helped distinguish between interventions where something of value was diverted from some students to be given to others (i.e., triage) and those where some students received something additional beyond what all students were offered.

The coders met weekly to discuss the types of interventions created by the schools and our interpretations of the codes and data. The use of emergent design (Strauss & Corbin, 1990) uncovered patterns in the way that schools organized their interventions not initially considered, with school plans differing across different dimensions. For example, every school offered small group instruction for their targeted students, and all but one hired substitute teachers to support that small group instruction. Furthermore, identifying triage in these schools was not as straightforward as anticipated when beginning the analysis, in part because many schools created multiple interventions.

During axial coding, I recoded the TAIPs to collect data on the dimensions which appeared to contribute to the equity and quality of the interventions. These features include which adult provided the intervention, which students were targeted, and when during the school day they occurred. These patterns ultimately led to grouping the interventions into four main types: (a) triage (teacher attention), (b) pullout tutoring, (c) substitute provided, and (d) support all students. To bolster the credibility of these groupings, I defined the features for each type of intervention—described in more detail in the Results section—and then used the school database to identify which schools included those features. I would then reread those schools' TAIP documents to see whether it reflected the broad intervention type identified. This process allowed me to clarify what I meant by each intervention type and then to refine that definition against the school proposals. I analyzed the TAIPs fully multiple times in response to (a) new information from additional TAIPs, (b) information from the district, and (c) further refinement of the intervention types.

During data analysis, I presented and discussed preliminary results with members of the district research team several times. These meetings allowed for checking my understanding of the district's programs and policies with what emerged from the document review. The resulting discussions provided context for the next iteration of analyses. These discussions also provided insight into how this work has implications for district leaders and the opportunity to further contextualize somewhat confusing results. One goal of presenting the preliminary results to the district and checking my interpretation of the results was to triangulate information across these modes of data collection to increase the trustworthiness of the results.

Results

Which Students Did School Leaders Plan to Target with Interventions?

This question was answered using the information from the TAIPs and triangulating that with the student quantitative data, with the results shown in Table 14. Schools indicated that on average 77 students would participate in the targeted interventions, ranging from 30 to 180 students per school. In total, the 29 schools planned for a total of approximately 2200 students to participate in the interventions.

Every single school targeted third- or seventh-grade students for interventions, as indicated by the "Intervention targets (Grade Level/Subject)" blank on each TAIP. This follows from the state's waiver accountability system which rated schools based on math and reading test performance for both (a) third and seventh grade students separately and (b) students in grades three through eight in aggregate. Despite the state's intentions that schools support the learning for all students, school leaders creating the TAIPs recognized that third- and seventh-grade students had a larger influence on their rating than students in other grades and offered interventions only for students in those grades. This is underscored in School A's TAIP (recreated in the top part of Figure 16) which states that "scores for 7th graders will count twice for overall school performance" before detailing that they plan to target seventh-grade Priority students. This indicates that School A's administrator was aware of and responsive to the metrics the state used in the waiver accountability system. School A is not alone; although this is the only school to explain in detail why this grade was selected, every middle school targeted seventh graders and every elementary school targeted third graders. Only one elementary school—3% of the sample—included students from another grade in their targeted interventions. This is important because it highlights the first equity issue of these interventions: only students from certain grades had access to additional remediation while other students were ignored.

Given that schools targeted third and seventh grade students, did they focus on Priority students? The results in Table 14 indicate that nearly one in three schools explicitly identified their targeted students as "Priority 1 and Priority 2" students, with twice as many middle schools than elementary schools using this description. It was harder to tell in other schools which third- and seventh-grade students were targeted. For example, 13 schools—including School B shown in the bottom panel of Figure 16—used the term "targeted students" and two additional schools used "identified students" to describe their focal students. Both of these terms come directly from the title of Figure 2, the district's *Identification of Target Students* document. This strongly suggests that these 15 schools were referring to Priority students, and one school made this clear by specifying that students were "grouped based on the targeted list (Priority 1 and 2)." This assumption was verified by comparing the number of students identified as Priority in the quantitative data with “the number of students participating” that was reported on the TAIP. Worth highlighting is that every school which used the “Priority” label included both Priority 1 and Priority 2 students in the interventions. There were no examples of schools that limited their target students only to those who were labeled Priority 1[19].

While most schools targeted students close to proficiency (i.e., bubble students), they did not necessarily use the Priority label to identify those students. For example, School A’s proposal explicitly describes their criteria for inclusion: seventh grade students who answered within three questions of the proficiency line on Benchmark B. Three other schools used the same metric to identify their target students[20]. In addition, about 20% of the TAIPs identified their target students as those who scored Basic on the benchmark exam, with five elementary schools and one middle school using this metric[21]. That elementary schools were more likely than middle schools to use the Basic label (which is from the benchmark) rather than the Priority label is likely related to the fact that third graders would not have prior year test scores. The Priority matrix from Figure 2 is based on students’ having a test history from the state. In cases where students are missing the state projection—the majority of whom are third-graders—the bottom line of Figure 2 indicates that schools should use the DEA benchmark score to identify their Priority students.

[19] This supports the decision to combine these two labels as a single “Priority” group in the Chapter 3 analyses.

[20] This is the same definition used in Alternative 2 of Chapter 3 to identify the bubble students in the years before the Priority label was available.

[21] This offers evidence that schools used the benchmark test-score labels to identify their bubble students, as hypothesized and tested in Chapter 2. Interestingly, these schools targeted only the students labeled Basic (ignoring students who scored just above proficiency on the benchmark). This is aligned with the Chapter 2 findings that elementary students barely labeled Basic gained significantly more than students barely labeled Proficiency.

Although schools used different ways to describe their bubble students (e.g., Basic, Priority, within three questions of proficiency, targeted, identified), all in all, 93% of TAIP schools did in fact focus on the students close to proficiency. While almost every school indicated that bubble students would receive the targeted interventions (the only exceptions were two proposals that did not include enough information to identify the targeted students), some schools expanded access to the interventions to other students. Nearly 25% of schools offered interventions to low-performers (described as "Below Basic" in two schools), with 17% of schools also supporting high performers. Elementary schools were more likely to include high- and low-performers than middle schools. This limited expansion of support, especially for low-achievers, highlights a second equity issue. Even though the state was under the waiver accountability system in this year, schools were still focused on increasing their proficiency rates—otherwise they would not target students close to that line—a direct result of state and district accountability systems which continued to use those metrics.

While the majority of the schools had a static targeted group for the entire intervention period, five schools explicitly stated that the groupings of students would be flexible. Schools that utilize flexible groups suggest a responsiveness to individual student needs. In some of these schools, students were grouped based on their individual gaps in understanding, and groups changed as the content of the intervention changed. For example, one middle school indicated that "teachers will meet each week to organize groups" and another noted that "after weekly probes, student groups will be restructured based on student need." Two schools indicated they would update their list of Priority students based on new data from the third benchmark exam. Middle schools were more likely to utilize flexible groups (29%) than elementary schools (13%).

The vast majority of the target student identification was based on student test scores. Only two schools mentioned any other criteria for inclusion. One school specifically targeted English Language Learners and another targeted students with disabilities. Both of these represent subgroups in the waiver accountability system used for calculating achievement gaps. That most schools selected students based on test scores, however, indicates that the TAIP schools seem more attuned to aggregate student performance than to individual subgroups.

To summarize, targeted students are comprised nearly entirely of third- and seventh-grade students. Of the 29 TAIP schools, 69% provided intervention to bubble students only, 7% supported bubble and low-performers, 17% offered remediation or enrichment for all students, and 7% were unclear about who was targeted. The school leaders typically used benchmark data shared by the district (including both the test-score and prescriptive labels investigated in Chapters 2 and 3). The description of the student selection process in a number of schools affirms some of the modeling decisions made in those chapters.

What Resources Were Used for the Targeted Interventions, and How Were the Interventions Structured?

Resources used. Most schools described their interventions without specifying the dollar amount required, but two schools provided a breakdown of how the TAIP funds would be used. These are illustrative of the types of resources that other schools mentioned in their proposals. One middle school asked for $15,680 to provide intervention for 160 seventh-grade students, with an average cost of about $98 per student. The intervention in that school would last for 30 days, and the money would be spent on (a) two teachers paid for two hours per day for 30 days for forfeiting their planning time to provide intervention ($3500), (b) six seventh-grade teachers paid to plan after school for one hour per day for 30 days ($5200), (c) one substitute teacher hired for 30 days ($3750), (d) two instructional coaches paid to plan for two hours per week for four weeks ($700), and (e) curricular supplies and paper ($2500).

An elementary school requested $8500 to target 45 third-grade students (averaging about $189 per student). That intervention would last 26 days and fund (a) one substitute teacher hired for 26 days to cover the class of the best reading and math teacher so that teacher could provide intervention to small groups of students ($5000[22]) and (b) two teachers paid to plan for two hours per day four days per week ($3500).

These types of expenses—hiring substitutes, paying teachers to plan after school, and purchasing intervention supplies—were common across the TAIPs. Table 15 shows what percentage of elementary and middle schools included each type of intervention feature, with the top part indicating how administrators proposed to spend the additional money. About a quarter of schools—five elementary and two middle schools—paid teachers to plan after school. This was done, as shown in the examples above, because teachers either (a) used their planning period to provide small group instruction to students or (b) needed additional time after school to plan the interventions. Approximately one-third of schools indicated that additional funding would go towards curricular materials (which includes both computer programs, such as Study Island, and workbooks focusing on state curriculum standards) or supplies, such as paper.

The most common expenditure across TAIPs was to hire substitute teachers. All but one school planned to use those funds to hire substitute teachers, and a large proportion of schools planned to hire two or more substitutes per day for multiple days. School leaders may already

[22] It is unclear why the per-day substitute costs from these two schools are not the same. The cost per day would be $125 for the middle school but $192 for the elementary school. Because substitute payment schedules are generally a set value across a district, it is unlikely that the daily cost would vary so much.

have had specific substitute teachers in mind to provide these interventions, such as retired teachers from the school. For example, one middle school specified the name of the substitute teacher they would hire for their interventions, and one elementary school indicated that their proposed substitute "has a teaching degree and previous experience with the [school's] student population." No other TAIPs indicated this level of detail regarding who would fill the substitute requests. In total, these 29 TAIP schools would require an additional 54 substitutes daily in the two months before the state test, above and beyond what is needed for normal teacher absences. This is an important feature to note because so many schools relied on substitutes to support small group instruction. If schools could not fill these requests, then their proposed interventions could not occur. This seems a concern in an elementary school which noted that they have "procured commitments from substitutes who regularly accept assignments." Prior research indicates that many schools have trouble getting substitutes to fill open jobs (e.g., Dorward, Hawkins, & Smith, 2000; Henderson, Protheroe, & Proch, 2002; National Education Association, 2001). The district did not share information on how frequently substitute jobs were filled during this time period, which is one reason why it is unknown how faithfully these plans were implemented.

Every proposal restructured the school day so targeted students would receive small group instruction in both math and reading. This small group instruction was provided by different people and at different times of day across the schools in this sample, as shown in the bottom part of Table 15. About one-quarter of the schools restructured their school days to provide multiple interventions, either to provide a different group of students with remediation or to focus on the target students in multiple ways (and at different times during the day). This is why some of the percentages shown in Table 15 add up to more than 100%.

The ways that schools utilized the substitutes was key for facilitating the small group instruction. Nearly a quarter of schools had substitutes directly teaching the targeted students in small groups, while others had the substitutes cover a teacher's class so the teacher could work with small groups of students. The teacher of record (meaning the targeted student's math or reading teacher) provided the intervention in the majority of schools (55%), although middle schools were more likely to do so (64%) than elementary schools (47%). One elementary principal wrote that "[o]ur belief is the 3rd grade teachers are the best equipped to provide targeted instruction to get the scores up," illustrating a focus on grade level expertise.

Schools used a variety of terms to describe the expertise or background of the teacher selected to provide the intervention. Terms included "expert," "best," "most effective," and "high performing." In some cases, schools provided no further clarification on how the school defined the term (e.g., School B from Figure 16). Other schools specified that teacher selection would be based on the state test performance of their previous students, their current students' benchmark scores, or their teaching evaluation scores. Nine schools indicated on the TAIP that expert teachers would provide the intervention with elementary schools using these educators more frequently than middle schools (40% compared to 21%). These differences make some sense given the context of elementary and middle schools. Elementary school teachers are responsible for teaching both reading and math content, but they may be better at teaching one subject than the other. Elementary principals may recognize the need to bring in content expertise from another grade to help fill in those gaps. Middle school teachers, on the other hand, are certified in their content area and may be better equipped to support their students' learning than another teacher.

Nearly two-thirds of schools indicated the time during the school day the intervention occurred. The most frequent time was during the core math or reading class, which was utilized in almost 60% of all plans but more often in middle (71%) than in elementary schools (47%). In four schools, the intervention occurred during another class, most likely an elective or related arts class. This indicates that the targeted students would be pulled out of these other classes to receive small group instruction *in addition to* their normal math and reading class. Five schools (three elementary and two middle) created or expanded a separate intervention period for students. In these cases, a period of time—ranging from 30 minutes to an hour—was set aside during the normal schedule for all students to receive instruction at their needed levels. This intervention time does not require students to be pulled from a normal class because all students are in this intervention period at the same time[23].

The district shared the *Identification of Target Students* document and blank TAIP proposals during a principals' meeting on February 6, 2013. The TAIP documents were approved by the district between February 12 and February 20, meaning that schools had fewer than two weeks to plan these interventions. The start dates of the interventions ranged from February 19 to March 4 while the end dates ranged between April 12 and April 23. The majority of schools began Monday, February 25 (N=13) and ended between April 19 and 23 (N=19). The number of days the intervention lasted varied across schools, as shown in Table 15. Sixty-two percent of schools held their interventions five days per week during this time period. Middle schools planned to have daily interventions more frequently than elementary schools (79% and 47%, respectively), whereas about 1/3 of elementary schools proposed to hold interventions only three days per week. Interventions would last on average 23.7 days in elementary schools and

[23] The way that these schools utilized the TAIP funding is explained in more detail in the next section.

29.0 days in middle schools. The total number of intervention days is shown in the boxplot in Figure 17. Elementary schools had more variation in the duration of their interventions, ranging from 9 to 33, with a median value of 24.5 days. Middle schools on the other hand, had a narrower range (from 23 to 40 days), with nine schools that varied only two days from the median of 29.5 days. There were approximately 150 school days in GCPS prior to the opening of the state test window, meaning the median intervention duration comprised 16.3% of the elementary school year and 19.7% of the middle school year. These interventions, then, represent a substantial amount of the school year prior to the state test.

Intervention structures. Looking at the individual features of the interventions reported in Tables 14 and 15, it is perhaps unsurprising that previous quantitative research has found mixed evidence of triage. The descriptive results point to several reasons for this and help guide the next analyses, which categorize the interventions into four different types. First, schools did not act like a monolith. The 29 schools combined intervention features in a number of ways, such that there was variation in the type, duration, provider, and focus of the interventions. The variation across interventions means that analyses which combine these schools together would not necessarily reveal test score differences between student groups. In the next part of this analysis, similar interventions were grouped together based on two dimensions: (a) equity and (b) quality. These dimensions, explained in more detail below, represent tradeoffs facing school leaders when designing the interventions, tradeoffs which have a direct effect on students and teachers.

The second reason for mixed evidence of triage may be that schools did not only target the bubble students. While the majority of TAIP schools planned to target only these students, a number of schools acted more equitably by supporting all students, most often during a built-in intervention period. If the schools that behaved more equitably are analyzed together with schools that engaged in triage, then differences in outcomes between bubble students and low-performers may not show up in aggregate test score comparisons. Equity of access was taken into account in two ways when grouping the TAIP interventions: (a) which students were targeted/had access to the interventions and (b) whether the resources provided to the targeted students were in addition to or at the expense of resources for students who did not have access to the interventions.

Third, the interventions varied in quality (i.e., their potential to provide increased learning opportunities for students). The quality of the intervention is influenced by the expertise of the individual who provides the intervention. Simply putting an adult with a small group of students would not necessarily lead to higher test scores. For instance, a number of schools planned for the substitute to provide the intervention. There are reasons to believe these individuals have less content knowledge expertise in math and reading to produce the additional learning that would translate into test score gains (e.g., Miller, Murnane, & Willett, 2008). On the other hand, both the teacher of record and an expert teacher would have the content knowledge expertise to work with struggling students, which should provide enhanced learning opportunities for targeted students.

Another aspect that influences students learning opportunities is the content and skills the interventions would focus on and the extent to which the intervention was aimed at meeting student needs. An intervention that focused on each individual's gaps in knowledge would likely provide increased learning opportunities. Some schools indicated their interventions would include a specific, targeted focus on individual student gaps. For instance, a middle school indicated that they had "looked at specific data on each student and know what each of them need intervention on to score proficient or advanced on [the state test].” In this school’s case, they focused on individual student needs. On the other hand, interventions where all students work on the same content rather than directly addressing student misconceptions might not increase student learning. For example, one school indicated that their intervention would focus on the “top three tested areas for [the state test].” An intervention with this one-size-fits-all approach may not provide enhanced learning opportunities because target students may be working on curricular standards that they already have mastered. Although the curricular focus is an important aspect of the intervention’s quality, few schools indicated the actual content focus beyond simply math and reading.

Intervention Type 1: Triage (teacher attention). One of the most common TAIP interventions proposed by schools leaders involved the teacher of record providing small group instruction for their own bubble students during normal math or reading class. These schools used the TAIP funds to hire substitute teachers who would stay with the rest of class while the teacher worked with their small group of students (e.g., School A from Figure 16). In some cases, the teacher taught all students during the first half of the period and then pulled the bubble students during the second half, leaving the substitute with the low- and high-performers. Other schools, however, had the teacher of record would work with the bubble students the entire class period while the substitute teacher taught the other students.

These interventions are labeled triage because the learning opportunities afforded to target students come at the expense of learning opportunities for non-targeted students. Bubble students get the attention of the teacher of record precisely because that teacher is not providing attention to other students. Even though these schools were given additional funds to target students, rather than providing additional resources to those students, they hired substitutes to divert resources (e.g., classroom teacher time) away from non-targeted students and towards targeted ones. One might expect that resources would need to be shifted in this fashion only in the absence of funding, yet a number of schools clearly used the additional funds to facilitate triage.

This use of additional funds was one of the most common responses across the TAIPs. In total, nine schools utilized this intervention, including three elementary and six middle schools. Schools planned to offer this intervention over a time period ranging from 12 to 31 days, with seven schools offering it for 24 or more days. This means that low-performing students in grades three and seven in some TAIP schools went more than five weeks before the state test without access to their normal math or reading teacher.

This intervention likely provides enhanced learning opportunities for targeted students because they receive additional attention from their teacher. Their teacher is likely well-positioned to support student learning because they would be a content area expert who knows their students' learning styles and individual gaps in knowledge. Some triage studies posit that "teacher attention" may be the mechanism through which triage occurs in the classroom (e.g., Krieg, 2008; Ladd & Lauen, 2010; Neal & Schanzenbach, 2010). This example of the teacher pulling a small group of students out for intensive tutoring represents one version of teacher attention.

Relating to equity of access, the non-targeted low- and high-performers left in the classroom with the substitute are likely to be negatively affected by this intervention. The substitute teacher may not have the math or reading content knowledge to help students who did not understand that day's lesson, something the teacher of record would do if she were in the class. Low-performing students might especially be harmed by this intervention because they clearly need additional support but would not necessarily have access to their teacher. This could help explain why the non-targeted (Below Basic) students' outcomes were worse when schools had TAIP funds than the following year (results found in the third alternative in Chapter 3).

This intervention creates some potential practical effects for students and teachers. For example, when targeted students are pulled to work with their teacher every day during class, they are publicly identified as receiving additional help. Non-targeted students might be able to place themselves in the low group, which risks de-motivating those students if they believe their teacher sees them as too low to do well. This public intervention could also have a positive effect on students who realize they are part of the high-performing students.

The movement of students and teacher for small group instruction disrupts the flow of class, but several schools limited the disruptiveness of this intervention. For instance, one middle school split the substitute between two teachers in order "to prevent a teacher from losing a complete instructional day with all students." Another middle school had their instructional coach remain in the class with the substitute when the teacher of record worked with the targeted bubble students. This would provide the students who were left behind access to a content area expert. Although modifications like these did not occur broadly, these schools acknowledged that losing access to their teacher might harm students and included in their plan ways to mitigate this issue.

There were two schools that had the teacher of record provide during-class interventions similar to triage, except the small groups were comprised of more than just bubble students. In these schools, the teachers would decide which students to pull for small group work, with one elementary school specifying that teachers would select the Basic and Below Basic students. The substitute still remained with the rest of the class, but these interventions were accessible to a broader group of students than those labeled as “triage.”

Intervention Type 2: Pullout tutoring. In the second intervention type, targeted students were pulled from a non-core class to receive small group instruction. Pullout tutoring generally occurred during a related arts class, such as art, band, or physical education. In schools that proposed pullout tutoring, all students were taught their normal math and reading class, but the targeted students— all of whom were bubble students—received additional tutoring at another time. Nine schools proposed pullout tutoring, including four elementary schools and five middle schools. Schools planned to offer this intervention over a time period ranging from 12 to 40 days.

TAIPs indicated that pullout tutoring would be administered by either (a) expert teachers in the school, (b) the teacher of record, or (c) substitute teachers. Because of their content knowledge, instruction offered by either the teacher of record or another expert teacher has the potential to provide students with increased learning opportunities. This is why I distinguish pullout tutoring from interventions provided by the substitute teacher, which are grouped separately and described next.

Overall, five schools indicated that the pullout would be administered by an expert teacher. When the “expert” or “effective” teacher provided the intervention during normal class time, the TAIP funds were used to hire a substitute who would cover this teacher’s class. As with triage, replacing a teacher with a substitute means that other students lose access to their teacher during their own math or reading time. School leaders did not specify whether these expert teachers come from another grade, but they likely do (given the focus on third- and seventh-grade students). Because many of the TAIP interventions would last more than five school weeks, this loss of time for students at other grade levels could affect their learning opportunities.

Several administrators limited how much time students lost with their math and reading teachers by using TAIP funds to pay teachers to plan their lessons after school. In two schools, teachers used their normal planning time, when they did not teach students, to pull small groups of students from other classes. When the teacher of record provides the intervention during her planning period, her other students do not lose access to her. The trade-off, however, is that the teacher—who generally has only one off period per day—works the entire school day with no break other than lunch and then stays after school to plan lessons. While teachers are being paid for their time, the extended school day without a break could contribute to teacher burnout. Two schools did not make it clear who would provide the pullout tutoring.

Because the tutoring would be provided by certified teachers, this intervention has the potential to provide enhanced learning opportunities. There are other factors that might influence the quality of the interventions. For example, pullout tutorials require lesson plans. The quality of those plans, and the extent to which they are focused on student learning gaps, would be key to increasing student learning. The TAIPs do not necessarily detail who would plan the intervention lessons, which limits how much can be said about their quality other than it is an important consideration for an effective intervention. The teachers would likely be able to provide individualized instruction to targeted students if they are provided with quality lesson plans.

Intervention Type 3: Substitute provided. The previous two intervention types were offered either during math or reading class (triage) or during another class (pullout tutoring). In both cases, math or reading teachers in the school provided the intervention, meaning the interventions were taught by content area experts. On the other hand, six schools (three elementary, three middle) proposed that substitutes would teach some or all of the intervention classes for targeted students. Three schools had the substitute provide small group instruction during students' normal math or reading class, while the other three had the substitute provide pullout tutoring during a non-core class. In most of these schools, this substitute-provided intervention occurred in concert with other intervention types, but two schools offered this as the only intervention available to bubble students. Several schools split the use of substitutes between freeing up teachers to offer the interventions and working with students in a computer lab. Across all schools, this intervention was offered to students ranging from 18 to 33 days, with five of the six schools proposing it to last 29 days or more. Using additional funds to have the substitute teacher provide the intervention represents an example of resources being provided to target students in addition to their normal support.

Substitute teachers may not have the content knowledge expertise, prior experience with the students, or knowledge of student misconceptions that would allow them to provide this type of targeted instruction (e.g., Clotfelter, Ladd, & Vigdor, 2009; Glatfelter, 2006; Gresham, Donihoo, & Cox, 2007). Being in a small group alone would not necessarily lead to improved learning if the substitute does not know the instructional materials or student needs. This leads to questions about who plans these interventions and the extent to which the interventions would offer additional learning opportunities for students. Some TAIPs detailed that teachers were paid to plan the intervention lessons after school, which implies that the lesson plans would be of quality and focused on student needs. It is unknown, however, how well substitutes would be able to implement these plans as written. Although some schools indicated that they would attempt to get a certified or specific teacher as substitutes, it is unclear whether that expertise was available at the district level. Substitutes, then, may be an inefficient use of resources because targeted students would not necessarily receive enhanced learning opportunities. Having substitutes monitor students in the computer lab, on the other hand, may be more beneficial for students because many computer programs target student needs.

When the substitute pulls students out of a non-core class for intervention, students do not lose time with their own teacher. The trade-off, however, is that the targeted student misses out on instruction from the non-core class. When the substitute provides intervention during normal class time, alternatively, time with the substitute replaces time with their teacher of record for targeted students. This is the opposite of the triage described previously. This substitute-provided intervention during core class actually frees up the teacher to work with low- and high-performers. This may actually benefit lower-performing students in the class because the class is smaller without the targeted students who are working with the substitute, meaning they could receive more attention from their own teacher during work time.

Intervention Type 4: Support all students. All three of the intervention types described above were, with few exceptions, focused only on bubble students and required those students to miss a normal class to receive the intervention. Five schools, including three elementary and two middle schools, proposed offering targeted instruction to all students through a built-in intervention time. Because all students participate in this built-in intervention period, schools proposed using a combination of the teachers of record, substitute teachers, instructional coaches, and other expert teachers to provide targeted instruction.

This intervention is in line with the Response to Intervention (RtI) model (Fuchs & Fuchs, 2006), a structure which provides increasing levels of supports for struggling students and involves flexible grouping and testing to ensure students are receiving targeted supports. The RtI model was originally intended as a method to identify students with disabilities but is often utilized school-wide to support the achievement of all students (Sailor, 2009). RtI can be built into school schedules by creating a separate intervention period during the school day for all students. During this intervention period, students who have gaps in their understanding of current content standards can receive remediation while students who are on grade level can work on enrichment or other activities. Schools typically utilize flexible groups during the intervention period so students who have similar gaps in their knowledge can be grouped together for targeted instruction based on those specific standards.

While most of the schools that offered this intervention already had this time built into their schedules, they used the TAIP funding for the same things already described: hiring substitutes, paying teachers to plan after school, and purchasing curriculum and supplies for the interventions[24]. In these schools, substitute teachers were used to provide instruction during the built-in intervention time, with one elementary school having the substitute teach "the largest enrichment group" (meaning the substitute worked with the high-performing students). Another elementary school specified that the substitute would be used "to make the groups smaller" during this time period.

Because the built-in intervention lasted for only one period during the school day, the substitute teachers hired with the TAIP money would have open time in their schedules. One school proposed that their substitute would plan and prepare lessons for the built-in intervention time (this is the school which specified they would hire a certified teacher who had experience with their student population, meaning the school leader believed the substitute had the expertise

[24] The middle school which detailed the cost of their intervention (described at the beginning of the results for RQ2) utilized this built-in intervention time.

to do this type of planning). Two of the schools used the TAIP funding to create more than one intervention, and they planned that the substitutes would provide some of the small group interventions just described—either during the normal math or reading class or during a non-core class—or overseeing the computer lab.

Some schools implementing this intervention used TAIP funds to pay teachers to plan after school, with an elementary principal noting that "teachers will need more planning time to prepare for the restructured school day." It makes sense that (a) flexibly grouping students and (b) targeting interventions at student need merits a substantial amount of planning. This work requires looking at student-level data to group students based on their gaps in learning and then deciding how to reteach concepts that these students did not understand the first time. To be done well, this focus on individual students requires time and data. Some TAIP schools recognized the needed time and paid teachers to do it after school.

Even within the "support all students" structure, there were examples of schools that provided bubble students with something in addition to or better than other students. For instance, two schools used TAIP funding to offer multiple interventions. This means that both (a) all students were supported and (b) bubble students were provided additional interventions such as triage or pullout tutoring. In another example, a middle school ensured that the bubble students would be taught by their teacher of record during the intervention period. This prioritization of bubble students indicates that the low-performing students would work with another teacher (who might not know them or their gaps very well) or with a substitute teacher (who may not have content knowledge). Similarly, one elementary school organized students into seven different groups based on their scores and made sure that the bubble students were placed into smaller groups than the other students. Both of these structures provide more individualized attention for bubble students. This is an interesting mix of providing additional or better supports for bubble students within this more equitable system.

For several reasons, the support all students intervention appears to represent the best practice of all the TAIP intervention types. First, it is the most equitable because the intervention was available to the broadest group of students. Low-performers, high-performers, and bubble students were all provided intervention. Second, all students had access to instruction at their needed level. Students struggling with the current content received needed remediation, while high performing students got to engage in enrichment activities. Although students across the test score distribution received support, these schools still focused TAIP funds only on third- and seventh-graders. While students from other grades were left out of the TAIP focus, schools that had this intervention built in to their schedules likely had it available for students in all grades. Third, this structure represents the least disruptive type of intervention because neither teachers nor students are missing out on anything to be in this intervention. It does not require students to be pulled from another class or to lose access to their teacher during normal class time. Given that most of the schools already had this structure built in to the schedule, students already had experience doing different things from their classmates, which limits the public identification of students who need additional help. Fourth, supporting all students is the least dependent intervention on substitute teachers or additional resources to facilitate the intervention. The triage or pullout tutoring would not occur if the substitutes did not show up, but the built in intervention time is flexible in regards to student group sizes and could be adjusted based on adult absences.

Discussion

The goal of this analysis was to get inside the black box and find out what schools planned to do with additional funds. The Targeted Academic Intervention Proposals represent the 29 school leaders' views about the best way to spend this money. A benefit of the TAIPs is that they offer a view into how a number of school leaders' designed interventions and decided which students should have access to them. Are these the type of school behaviors that were intended by state accountability policy and district data sharing initiatives? The interventions described in Chapter 4 have implications for state and district policymakers and should broaden their conversations about the effect of policy on students.

Takeaway 1: Incentives Matter

It is clear from the state's waiver application that they desired equity, using the phrases "every student" and "all students" dozens of times in that document. State policymakers intend for all students to improve over their baseline results every year. The design of the incentives under the waiver, however, did not necessarily align with those intentions and created competing incentives for districts, school leaders, and teachers. Furthermore, despite speaking in depth about equity concerns regarding schools who narrowly target bubble students, GCPS provided the data and encouraged schools to focus on bubble students before the state test. State policy influenced the district rules governing their TAIP program, which affected equity of access to support and resources because schools selected students for interventions based on state incentives.

The use of aggregate proficiency rates for students in grades three through eight did not "mitigate an overemphasis on grades three and seven" (ESEA Flexibility Request, 2012, p. 42). The state recognized that measuring schools based on single grades might lead schools to behave strategically, including in the waiver application that "[u]nderstanding that [annual measureable objectives] drive behavior, we added aggregate grades 3-8 Math and Reading measures to mitigate an over-emphasis on 3rd and 7th grade" (ESEA Flexibility

Request, 2012, p. 42). While recognizing that measuring school performance on individual grades might lead schools to focus on students in those grades, policymakers' fix—including the aggregate score—did not work. Every TAIP school targeted third or seventh grade students with interventions. This emphasis on individual grades appears to be driven directly by the state accountability system because none of the district's communication to the schools included mention of which grade to target. An important takeaway for state and federal policymakers is ensuring that the policy metrics and incentives match their intentions. This state appears to have learned its lesson, including in the subsequent accountability waiver application that "based on feedback gathered from stakeholders, measures that focus on individual grades (i.e., 3rd and 7th grade) are eliminated" (ESEA Flexibility Request, 2015, p. 45).

Proficiency rates caused schools to focus on bubble students. While the district encouraged schools to target Priority students, school leaders were allowed to determine the focal students for interventions. The Priority label alone represents a strong signal to schools that these are the important students, and 93% of schools followed this signal by targeting bubble students. Even though some of these schools also supported other students, nearly seven in ten schools focused interventions solely on students close to proficiency.

This focus on bubble students reflects both district and state accountability incentives that rate districts, schools, administrators, and teachers (and students), at least in part, based on proficiency rates. The state used proficiency rates in the waiver accountability system to (a) hold districts accountable, (b) identify the lowest-performing 5% of schools statewide, and (c) comprise 15% of principal and teacher evaluation scores. The waiver shifted individual school accountability from the state to districts, and GCPS' school accountability framework included multiple indicators, with 15% of school ratings determined by the school's proficiency rate.

The focus on bubble students is understandable, given the design of the accountability incentives. For example, because the state held districts accountable for proficiency, it makes sense that district leaders would want schools to know who their bubble students were so they could focus on students close to that line. In addition, the lowest-performing 5% of schools risked possible takeover by the state, a possibility determined solely by proficiency rates. Given that the schools with TAIP funding were amongst the lowest-performing in the district, it makes sense that these schools would want to increase this metric to avoid being in the bottom 5% of schools in the state. Despite including student growth metrics in the policies, the use of proficiency rates in both state and district accountability appears to have led directly to the TAIP program in GCPS where schools focused on bubble students. While I cannot extrapolate these behaviors beyond GCPS, or even beyond the schools that received TAIP funding, it is not difficult to imagine that more schools across the state felt that their rating would benefit through this type of focus.

Students in need are left behind. Schools did not simply focus on bubble students, however, but on bubble students in grades that counted twice. Micro-targeting third- and seventh-grade bubble students leads to equity concerns, especially regarding the students who do not get access to needed supports and are left behind. For example, in all of the TAIPs, because the schools targeted their interventions only for third- and seventh-graders, students in other grades may not receive needed additional support.

Only one in four TAIP schools expanded their TAIP interventions to include low-performers, with even fewer providing enrichment for high-performers. The district labeled the lowest-performing students as needing a “Multi-Year Plan” to meet proficiency. Yet the district did not require that schools turn in a plan for those students, only a proposal for how schools would target Priority students. By not asking schools to detail how they will improve outcomes for the low-performing students, and by labeling these students as “Multi-Year Plan” and color-coding them red in Figure 2, the district implicitly wrote off those students. Low-performers being neglected is a major concern regarding educational triage and has equity implications in this district. In order to get additional resources to support their learning, students needed to be in an important grade and to score right around proficiency on the benchmark. This is concerning because schools appear to be making instructional decisions based on the metrics used to rate them and not necessarily on student need.

The limited expansion of support to other students is at odds with a stated goal in the waiver, which describes a “focus on growing every student, every year” (ESEA Flexibility Request, 2012, p. 43). The state designed the evaluation system to incent individual teachers and principals to focus on student growth by making 35% of teacher and principal evaluation scores based on value-add metrics. That only a handful of TAIP schools targeted students across the test score distribution, however, suggests that the incentives regarding proficiency rates might be stronger than those regarding growth metrics.

The TAIP analysis indicates that incentives matter. There appear to be multiple external pressures that led many schools to micro-target bubble students in third- and seventh-grade. This behavior suggests a narrow focus on the measured aspects of schools, as Campbell (1979) would have predicted. The TAIP proposals reflect the language and metrics from both the district and the state, which points to how carefully members of these role groups should consider the messages they send to those who implement the policy at the school level. This study indicates that policymakers can affect equity of access to enhanced learning opportunities through the incentives and metrics in accountability systems.

Takeaway 2: Triage Occurs Even When Additional Resources Are Available

While schools might not have been able to implement the proposals exactly as written, it is clear that a large number of schools planned to engage in triage. Nine schools proposed using TAIP funding to hire substitute teachers who would stay with low- and high-performers while the teacher of record would work with small groups of bubble students during class. That these schools used additional resources to facilitate the diversion of resources is counterintuitive, given that the additional funds should have alleviated the need to shuffle resources.

Furthermore, the loss of time with one's teacher can have serious consequences. Miller, Murnane, and Willett (2008) use three years of data from a large urban district and estimate that fourth grade math achievement dropped by 3.2% of a standard deviation for every 10 days a teacher is absent. Some of the proposed interventions would last for up to three times that length, suggesting that students who were left behind lost something of real value.

As just noted, nearly 70% of schools planned to focus solely on bubble students, and nearly half of those proposed to do so through triage (i.e., diverting resources to targeted students). The other schools focused on bubble students either through pullouts or by remediation with a substitute (i.e., providing something additional to targeted students). Because these pullout interventions classes occur in addition to normal classes, this intervention does not offer targeted students resources at the expense of non-targeted students. On the surface, that makes this practice seem more equitable to the non-targeted students because they do not lose out on their teacher during class time (like in triage), they are simply left out of receiving additional support[25]. But if students do not get additional remediation because they are viewed as

[25] Although if the teacher is pulled from another grade, the students who are left with a substitute are certainly affected.

too low, then it is worth considering whether this allocation of resources is really more equitable or whether it is somehow just more palatable than directly diverting resources from low-achieving students.

Takeaway 3: Not All Interventions Are Equal

School leaders used a variety of strategies to restructure their school day, each representing tradeoffs related to student learning opportunities, equity of access, and disruption of normal schedules. The most common interventions were teacher attention, or triage. This intervention represents real equity issues because the low-performers lose access to their teacher for some or all of their math or reading class. The second intervention type was pullout tutoring, where students receive an additional remediation class in addition to normal math and reading classes. The quality of this intervention relies on who provides it (specifically the content knowledge expertise of the provider) and the curricular focus. The third type of intervention had the substitute teachers provide the intervention. There appears to be an equity versus efficiency tradeoff happening in the use of substitutes. Having substitutes work with small groups of students would not necessarily increase learning and would be an inefficient use of resources if the subs (a) are not content area specialists and/or (b) do not know the students gaps or needs. The equity tradeoff, however, is that using substitutes to offer remediation limits how much time other students would lose with their teacher. Each of the TAIP interventions provide targeted instruction to only a small group of students. Schools can decide which students can access those resources, but each of them has limited space so some students have to be left out.

The last type of intervention—support all students—appears to represent a best practice for schools. It seems the most equitable because it provides support for students across the test store distribution. The curriculum and student groupings are based on current student needs. It is the least dependent on substitute teachers. Because the intervention time is built into the school day, neither students nor teachers have to adjust their normal schedules. This combination of factors makes the “support all students” intervention align closest with the intentions of the district leaders and state policymakers.

Despite these benefits, it is not easy to implement a high-quality RtI model school-wide (e.g., Burns & Gibbons, 2012). The RtI model requires teacher to use student data to flexibly group students by need and to plan remediation for that shared gap in knowledge. Schools need data that identifies what students know and on what they need additional help, but using data to support students in need is a challenge for administrators and teachers (Means, Padilla, Gallagher, & SRI International, 2010). Beyond creating a school schedule with built-in time for enrichment and remediation, leaders and teachers need (a) training to effectively use the student data and (b) time to look at the data so they can determine which students need what support. The actual lesson structure and activities need to be of quality, and GCPS schools likely face challenges in expertise and time to plan and create quality lessons that address student misconceptions.

Despite these challenges, the support all students structure is focused on student growth for all students—including high-performers, who are often left out of the conversation. Policymakers at the state and district levels express a clear focus on growth for all students. Even though this structure is difficult to implement well, considering how to support schools to effectively use student data focuses discussions on student learning. This suggests an opportunity for additional research to explore how educators could implement these school-wide interventions more effectively.

All schools planned to provide targeted students with small group instruction, and the main use of TAIP funding was to hire substitute teachers to facilitate small group instruction, a finding not anticipated at the beginning of the study. This highlights a possible source of expertise for districts and schools. Districts might consider training a cadre of substitutes to provide this type of targeted intervention. Retired teachers, for example, could be used as interventionists for the short period of time before the state test. Additional research on how to best utilize substitute teacher would benefit schools looking to provide targeted instruction to students.

The variation in the quality of the TAIP proposals indicates that these schools were not necessarily able to design quality interventions given the short time frame. Given that identifying the target students, designing the interventions, turning in the TAIP documents, and receiving approval for these interventions all occurred in fewer than two weeks, it is not clear what process the district used to mark the TAIPs as “Approved.” For example, it is unclear the extent to which district leaders considered whether these interventions (a) would provide additional learning opportunities for students or (b) were widely accessible for struggling students. Thinking about these factors seems worthwhile for districts to do before they offer funding under programs such as the TAIP interventions. This implies that districts should provide schools more time and more guidance to design interventions which better reflect district intentions. This additional guidance would require district leaders to discuss what quality interventions would look like, which would focus discussions on student learning.

Given the results from this analysis, I offer some questions that district and school leaders might consider when discussing interventions which occur during the school day. The first consideration relates to equity of access. Which students need additional support and on what content? Is selection into intervention based on student needs or other concerns? What is being lost and for whom? The second consideration relates to quality. How tailored are the interventions to student needs? If students are not provided targeted support aimed at their gaps in learning, they will not have the opportunity for increased learning. This type of targeting requires careful preparation and planning of the lesson to ensure it fills in gaps in learning. Who is going to prepare these interventions? And will the person who prepares the lessons also be providing the intervention? If not, how much preparation time does the provider have? Schools need access to quality planners and providers if they are expected to support student learning.

Limitations

This analysis focused on how school leaders planned to spend TAIP funds. It does not represent what these TAIP schools were doing for the rest of the school year to support students. While these results indicate that schools behaved strategically by targeting interventions mainly for bubble students, with many schools doing so by diverting resources from low-performers, this does not mean that schools behaved this way for the entire school year.

In addition, as noted already, these documents are proposals for how schools planned to structure their interventions. We do not know how faithfully schools were able to implement these proposals. That limits the ability to quantify the effects of the various interventions on student outcomes.

Tables

Table 13. School summary statistics for schools with and without TAIPs

		Elementary Schools		Middle Schools	
		Has TAIP	No Plan	Has TAIP	No Plan
Math	*Prior Year Test Label*				
	Proficient/Advanced	36.1	40.1	36.8	38.3
		(8.3)	(17.6)	(12.1)	(21.1)
	Basic	49.1	47.1	37.1	36.1
		(5.1)	(12.1)	(2.8)	(9.5)
	Below Basic	14.8	12.8	26.1	25.6
		(4.6)	(8.9)	(11.5)	(14.4)
	Prescriptive Label				
	Enrichment	14.7	19.1	11.1	16.1
		(35.4)	(39.3)	(31.4)	(36.7)
	Priority	44.9	43.8	40.1	38.1
		(49.7)	(49.6)	(49.1)	(48.6)
	Multi-Year Plan	40.4	37.1	48.0	45.8
		(49.1)	(48.3)	(50.0)	(49.8)
Reading	*Prior Year Test Label*				
	Proficient/Advanced	34.1	40.2	40.1	40.0
		(8.2)	(18.7)	(10.0)	(22.0)
	Basic	50.9	47.1	46.0	44.4
		(3.7)	(12.6)	(5.6)	(9.3)
	Below Basic	15.0	12.7	13.9	18.2
		(5.4)	(8.0)	(7.2)	(10.7)
	Prescriptive Label				
	Enrichment	15.0	21.5	16.6	22.9
		(35.7)	(41.1)	(37.2)	(42.0)
	Priority	37.2	35.5	40.3	35.1
		(48.3)	(47.8)	(49.1)	(47.7)
	Multi-Year Plan	47.8	43.0	43.1	41.9
		(50.0)	(49.5)	(49.5)	(49.3)
Student Demographics	Black	29.5	48.8	42.2	55.5
		(11.5)	(31.4)	(20.5)	(22.2)
	Hispanic	33.2	13.5	18.2	10.8
		(18.0)	(15.9)	(13.4)	(12.4)
	White	32.2	33.4	34.6	30.5
		(13.2)	(26.6)	(16.6)	(18.5)
	Other	4.0	2.9	4.3	2.7
		(2.9)	(3.5)	(4.9)	(2.4)
	Economically disadvantaged	82.7	75.3	76.8	77.6
		(14.8)	(25.9)	(11.1)	(22.2)
	English language learners	37.1	13.9	10.9	8.9
		(19.7)	(15.6)	(11.6)	(10.3)
	Students with disabilities	11.3	13.2	12.8	14.0
		(2.4)	(3.5)	(3.0)	(5.5)
	Mobility rate	28.9	29.9	27.9	32.7
		(7.1)	(14.3)	(12.7)	(19.0)
	Observations	15	59	14	22

Standard deviations in parentheses

Table 14. Descriptive statistics of which students were targeted with TAIP interventions

	Total	Elementary	Middle
Number of students targeted	Average: 76.8 Range: 30-180	Average: 61.9 Range: 30-180	Average: 91.2 Range: 47-160
3rd or 7th grade	29 (100%)	15 (100%)	14 (100%)
Other grade included	1 (3%)	1 (7%)	0 (0%)
Bubble students	27 (93%)	14 (93%)	13 (93%)
Priority 1 and 2	9 (31%)	3 (20%)	6 (43%)
Close to proficiency ("within 3 questions of proficiency")	4 (14%)	2 (13%)	2 (14%)
Basic	6 (21%)	5 (33%)	1 (7%)
"Targeted students"	13 (45%)	8 (53%)	5 (36%)
"Identified students"	2 (7%)	0	2 (14%)
Low performers (Below Basic, remediation)	7 (24%)	4 (27%)	3 (21%)
High performers (enrichment)	5 (17%)	3 (20%)	2 (14%)
Bubble students only	20 (69%)	10 (67%)	10 (71%)
Bubble and low performers students	2 (7%)	1 (7%)	1 (7%)
All students receive intervention or enrichment	5 (17%)	3 (20%)	2 (14%)
Unclear who was targeted	2 (7%)	1 (7%)	1 (7%)
Flexible groups	5 (17%)	2 (13%)	3 (29%)
Subgroups (e.g., ELLs, students with disabilities)	2 (7%)	1 (7%)	1 (7%)
Observations	29	15	14

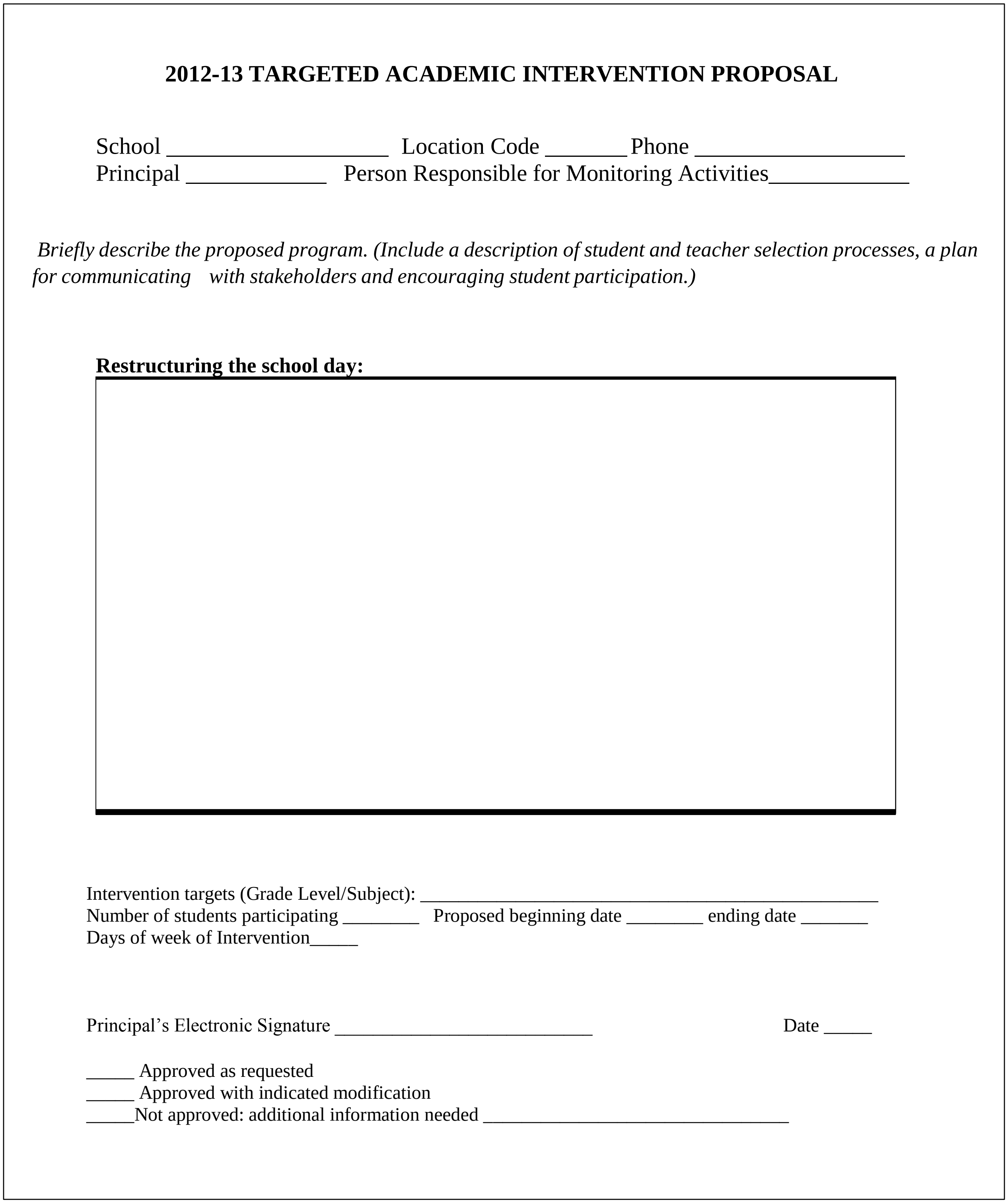
Table 15. Descriptive statistics of TAIP intervention features

	Total	Elementary	Middle
How were additional resources spent?			
Hire subs	28 (97%)	14 (93%)	14 (100%)
Curricular materials and supplies (e.g., student workbooks, computer programs, paper, copy toner)	9 (31%)	4 (27%)	5 (36%)
Teachers paid to plan after school	7 (24%)	5 (33%)	2 (14%)
Small group instruction	29 (100%)	15 (100%)	14 (100%)
Multiple interventions	7 (24%)	3 (20%)	4 (29%)
*Who provided the intervention?**			
Teacher of record	16 (55%)	7 (47%)	9 (64%)
Other teacher (e.g., “expert,” “most effective”)	9 (31%)	6 (40%)	3 (21%)
Substitute	7 (24%)	4 (27%)	3 (21%)
Unclear who provided	3 (10%)	2 (13%)	1 (7%)
*When during the school day did the intervention occur?**			
During core math or reading class	17 (59%)	7 (47%)	10 (71%)
During another class	4 (14%)	1 (7%)	3 (21%)
Expanded or created separate intervention time	5 (17%)	3 (20%)	2 (14%)
Unclear when during the day	10 (34%)	7 (47%)	3 (21%)
How frequently did the intervention occur?			
2 days per week	2 (7%)	1 (7%)	1 (7%)
3 days per week	5 (17%)	5 (33%)	0 (0%)
4 days per week	4 (14%)	2 (13%)	2 (14%)
5 days per week	18 (62%)	7 (47%)	11 (79%)
Total number of days	26.2	23.7	29.0
What was the curricular focus?			
Math	29 (100%)	15 (100%)	14 (100%)
Reading	29 (100%)	15 (100%)	14 (100%)
Observations	29	15	14

*may not add up to 100% because schools could have multiple in this category

Figures

Figure 15. Blank Targeted Academic Intervention Proposal (TAIP)



2012-13 TARGETED ACADEMIC INTERVENTION PROPOSAL

School ___________________ Location Code _______ Phone __________________
Principal ____________ Person Responsible for Monitoring Activities____________

Briefly describe the proposed program. (Include a description of student and teacher selection processes, a plan for communicating with stakeholders and encouraging student participation.)

Restructuring the school day:

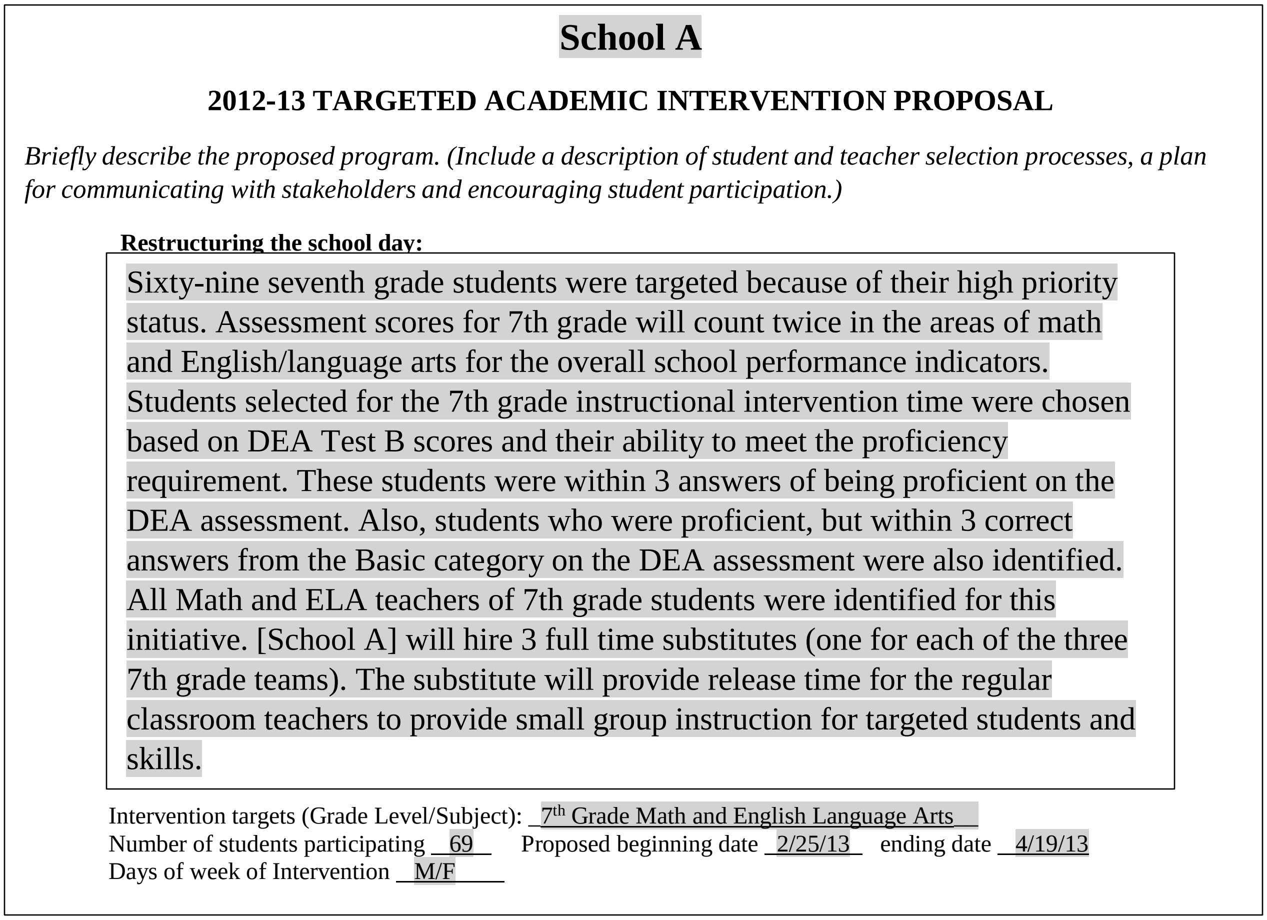
Intervention targets (Grade Level/Subject): _______________________________________
Number of students participating ________ Proposed beginning date ________ ending date _______
Days of week of Intervention_____

Principal's Electronic Signature ___________________________ Date _____

_____ Approved as requested
_____ Approved with indicated modification
_____Not approved: additional information needed ________________________________

Figure 16. Samples of completed TAIPs

Panel A. School A



School A

2012-13 TARGETED ACADEMIC INTERVENTION PROPOSAL

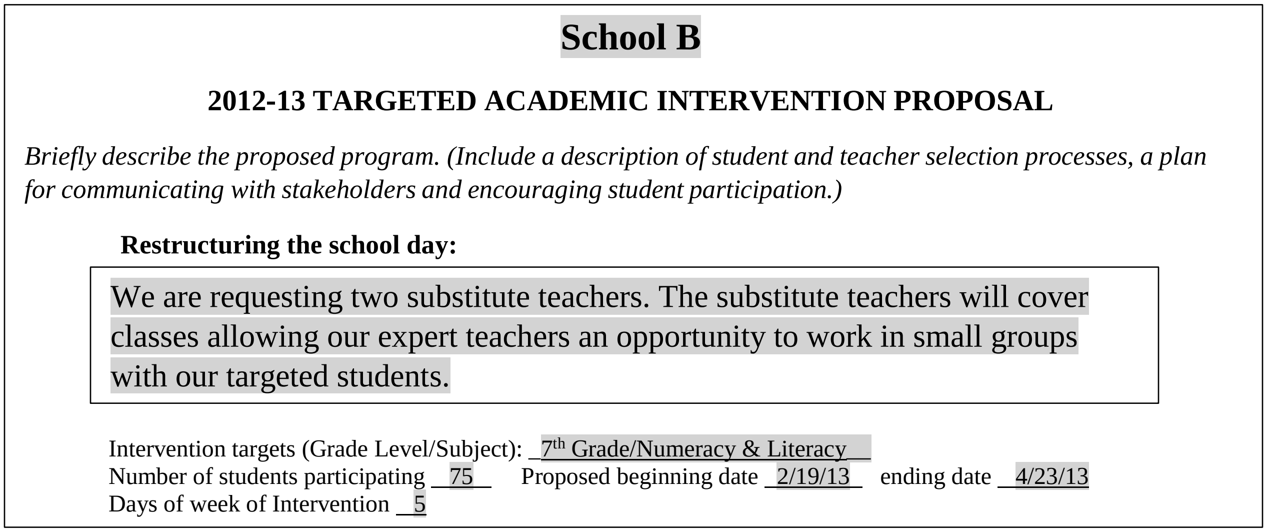
Briefly describe the proposed program. (Include a description of student and teacher selection processes, a plan for communicating with stakeholders and encouraging student participation.)

Restructuring the school day:

Sixty-nine seventh grade students were targeted because of their high priority status. Assessment scores for 7th grade will count twice in the areas of math and English/language arts for the overall school performance indicators. Students selected for the 7th grade instructional intervention time were chosen based on DEA Test B scores and their ability to meet the proficiency requirement. These students were within 3 answers of being proficient on the DEA assessment. Also, students who were proficient, but within 3 correct answers from the Basic category on the DEA assessment were also identified. All Math and ELA teachers of 7th grade students were identified for this initiative. [School A] will hire 3 full time substitutes (one for each of the three 7th grade teams). The substitute will provide release time for the regular classroom teachers to provide small group instruction for targeted students and skills.

Intervention targets (Grade Level/Subject): _7th Grade Math and English Language Arts__
Number of students participating __69__ Proposed beginning date _2/25/13_ ending date __4/19/13
Days of week of Intervention __M/F____

Panel B. School B



School B

2012-13 TARGETED ACADEMIC INTERVENTION PROPOSAL

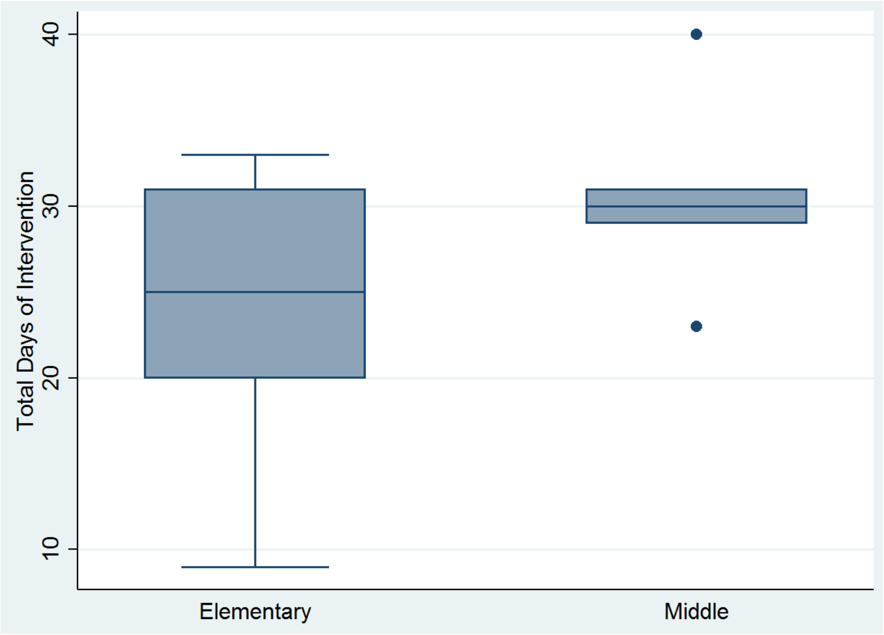
Briefly describe the proposed program. (Include a description of student and teacher selection processes, a plan for communicating with stakeholders and encouraging student participation.)

Restructuring the school day:

We are requesting two substitute teachers. The substitute teachers will cover classes allowing our expert teachers an opportunity to work in small groups with our targeted students.

Intervention targets (Grade Level/Subject): _7th Grade/Numeracy & Literacy__
Number of students participating __75__ Proposed beginning date _2/19/13_ ending date __4/23/13
Days of week of Intervention __5

Figure 17. Range in total number of days that proposed interventions would last



Chapter 5: Conclusion

The purpose of this dissertation was to explore how schools in this urban district used the benchmark data that they received three times per year and to examine the extent to which schools responded to the changing incentives of state policy. Each analysis chapter included a thorough discussion of the results. Here, I recap the results from the analyses before discussing the broad takeaways found across the chapters.

Summary of Results

Chapter 2 used five years of benchmark data to investigate whether schools used benchmark information differently over the course of the school year and the extent to which that changed when the policy incentives changed. That analysis exploited the fact that students were assigned test-score labels of Below Basic, Basic, Proficient, and Advanced based on the number of questions answered correctly. Local linear regression discontinuity methods were used to estimate the causal effect of receiving the higher of two labels during NCLB, and the RD models were modified to utilize difference-in-differences methods to quantify the differential effect of the labels during the waiver.

The results from Chapter 2 indicated that schools did focus on bubble students during NCLB. The focus on bubble students occurred more strongly in math than in reading, and it occurred only after the third benchmark. Students barely labeled as Below Basic scored about 0.04 standard deviations lower than students barely labeled Basic on the third math benchmark during NCLB (estimates which were consistently significant). In reading, there was some evidence that middle school students labeled Below Basic scored 0.02 to 0.03 standard deviations lower than students barely labeled Basic on the third reading benchmark. The reading estimates from NCLB were stable but significant only at higher bandwidths. Elementary and middle schools both focused on Basic students during NCLB, but they did so at different thresholds. Middle schools were more consistent across subjects with their behavior during NCLB, focusing on Basic over Below Basic students in both math and reading. Elementary schools, on the other hand, focused on Basic over Proficient students in math only.

While Chapter 2 found that GCPS schools targeted bubble students during NCLB, the adoption of the waiver accountability system counteracted that effect, implying that schools treated lower-performing students better during the waiver. Students labeled as Below Basic on Benchmark C benefited in both subjects and in both types of schools. The discontinuity shifted 0.06 standard deviations towards students barely labeled Below Basic in math, and 0.04 standard deviations in reading. The shift in focus in math occurred earlier in the year (by Benchmark B) and at more thresholds (both Basic and Proficient threshold) than in reading. Middle schools were again more consistent in their shift during the waiver across subjects.

These effect sizes of 0.02 to 0.06, while statistically significant, are relatively small in magnitude. Although other triage studies have found larger effect sizes, these estimates are within the lower end of the range of those studies. For example, Jennings and Sohn (2014) find that the lowest-performing students in Houston public schools lost about 0.11 standard deviations (SD) in math after NCLB was introduced. Neal and Schanzenbach (2010) estimate that students in the lowest deciles scored about 0.04 SD worse while students in the middle deciles scored about 0.13 SD better. My estimates are similar to those found by Krieg (2008), that bubble students performed between 0.02 and 0.05 SD better in schools facing accountability pressure during NCLB.

Chapter 3 investigated alternative explanations for the Chapter 2 results. The first alternative explored whether the inclusion of growth metrics in the educator evaluation system contributed to the previous chapter's findings. This analysis used the natural experiment that occurred because the new evaluation system was adopted in 2011-12, the year before the waiver was implemented. These results showed that the shift towards lower-performing students began after the new evaluation system—but before the waiver—was adopted. That shift was not as broad as the previous chapter. The results indicate that the new evaluation system led to a shift in focus (a) only at the Basic threshold, (b) only on Benchmark C, and (c) more consistently in math than in reading. Students who barely scored Below Basic on math Benchmark C gained about 0.04 standard deviations more in 2011-12 than similar students in previous years (estimates which are statistically significant except at the narrowest bandwidth). These results are similar to those from the previous chapter, although the estimated differential effect of the new evaluation system is slightly smaller in magnitude. In reading, students who barely scored Below Basic on the third benchmark gained between 0.010 and 0.065, estimates which get larger in magnitude and are significant only at the wider bandwidths.

The second alternative investigated in Chapter 3 was that schools continued to triage during the waiver but used a different definition of bubble students than the Basic label that was explored in Chapter 2. This analysis took advantage of a set of prescriptive labels supplied by the district beginning in 2012-13 which explicitly identified the students close to proficiency as Priority, with low- and high-performers labeled as Multi-Year Plan and Enrichment, respectively. Using difference-in-differences models, this examination compared outcomes for students in these groups during NCLB and during the waiver.

The results from this alternative affirm the previous chapter's results in math that schools focused on bubble students during NCLB and that low-performers benefitted after the waiver. As with the majority of the results, these findings were concerntrated on the third benchmark. In math, bubble students gained 0.031 standard deviations more than low-achievers during NCLB, and there was a shift during the waiver where low-performers gained 0.043 standard deviations more. These results are similar in magnitude to the estimates at the Basic threshold from Chapter 2. In addition, high-achievers gained significantly less than bubble students (0.021 standard deviations) during NCLB, indicating that the gains for bubble students in math came at the expense of low- and high-performers (i.e., schools engaged in triage). The reading results differed from those in Chapter 2, with low-performing students on the third reading benchmark gaining 0.033 standard deviations more than bubble students during NCLB.

The third alternative explored in Chapter 3 is that the reduced focus on bubble students occurred because 29 schools received additional funding in 2012-13 to provide students with targeted academic interventions. Under this hypothesis, the TAIP funding allowed schools to target bubble students without having to divert resources from low- and high-achievers. To test this alternative, the models from Chapter 2 were modified to include variables indicating whether (a) students attended a school that would receive TAIP funding and (b) the year was 2012-13, when the funding was provided to schools.

The results of this analysis indicate that the TAIP funding does not explain the reduced focus on Basic students found during the waiver. The gains for low-performers persisted after controlling for TAIP money and year. The coefficients on the difference-in-difference-in-differences estimator for students labeled Basic on the third benchmark in TAIP-funded schools during 2012-13 were consistently positive in both reading and math (although they were only rarely statistically significant). Positive coefficients indicate that students who were assigned the Basic label gained more than (a) students assigned the Below Basic label in the same year and (b) students assigned the Basic label the following year. That is, the TAIP money benefitted the

Basic students. The most surprising result from this analysis was that the non-targeted Below Basic students did worse when their schools had TAIP funds to design interventions. This suggested that schools used the TAIP funds in ways that harmed the low-performers, and Chapter 4 confirmed that many schools used the TAIP funds to facilitate triage.

Chapter 4 analyzed the Targeted Academic Intervention Proposals completed by 29 school leaders that described how schools planned to spend the additional funds and which students would be targeted. Every school leader proposed to offer interventions only for third and seventh graders (the grades which counted twice in the waiver accountability system), with nearly 70% of the schools micro-targeting interventions for bubble students in those grades. Only a quarter of schools included low-performers in these interventions, which helps explain why the TAIP funding was not related to a reduced focus on bubble students (i.e., the third alternative in Chapter 3).

Chapter 4 found that TAIP funding would be spent primarily on hiring substitute teachers, paying teachers to plan after schools, and purchasing curriculum and supplies. The interventions were grouped into four types based on when during the school day the intervention occurred, who taught the intervention, and which students had access to the interventions. Even though schools were given additional funds, the most common use of TAIP funding was to facilitate triage (i.e., teacher attention), where substitutes worked with low- and high-performers while the teacher of record worked with bubble students. A large number of schools also offered bubble students pull-out tutoring, where an expert teacher provided small group instruction during a non-core class. Some schools used the substitutes to provide the intervention. The most equitable intervention structure was found in schools that built in an intervention period to the school day and had all students receiving instruction at their needed levels.

Takeaways

The takeaways from this dissertation are as follows.

1. Schools in GCPS did focus on bubble students during NCLB to the detriment of low-performers. This focus occurred after the third benchmark (around February of the school year). Schools used benchmark data to facilitate this focus, including the test-score labels and the prescriptive labels shared by the district.
2. The quantitative analyses indicate that schools responded to changes in policy incentives by shifting focus to lower-performing students. Those policy changes included both school accountability policy and teacher and principal evaluation policy.
3. The results across the quantitative analyses were generally more consistent and stronger in math than in reading.
4. Middle and elementary schools behaved differently. Middle schools appeared to be more responsive to accountability incentives, having more consistent results across subjects and accountability eras.
5. Even though the quantitative analyses show that, on average across the district, lower-performing students gained when the incentives changed, school documents indicate that the schools which received additional funding in 2012-13 focused interventions before the state test on bubble students. Even though they had additional funds to provide these targeted interventions, many schools diverted resources from low- and high-performers to offer these interventions (i.e., engaged in triage).

Implications

Researchers

This dissertation has implications for researchers, especially those who are engaging in triage research. When the quantitative data were broken down by TAIP and by year (i.e., in Chapter 3 Alternative 3), TAIP schools' interventions—which targeted bubble students but not low-performers–were reflected in the positive DIDID estimates for Basic students. Furthermore, the use of TAIP funds to engage in triage was also reflected in the negative estimates for Below Basic students in 2012-13 compared to the 2013-14. This evidence of triage was visible only because of (a) access to district-level data (including benchmark scores and school documents) and (b) deep contextual knowledge of the district's programs. These behaviors were based on students' benchmarks scores, something that previous researchers studying triage have not neccesarily been able to access. That means these behaviors would be hidden from researchers and policymakers. Relying on test score differences to indicate that schools are behaving inequitably is concerning if scores do not pick up on these hidden behaviors.

Because the district-wide test scores do not reflect this focus on bubble students, this has implications for researchers who use test scores as proxies for strategic behavior (e.g., Springer, 2008b). In the Introduction, I offered as one of several hypotheses that prior triage research has failed to detect a focus on bubble students because researchers may inaccurately identify the bubble students. This hypothesis is borne out by this project. I illustrate this by comparing the district-supplied Priority label (which explicitly identifies GCPS' bubble students) with two common methods prior researchers have used to identify bubble students: (a) dividing the prior year test score distribution into quantiles, with the middle quantiles representing the bubble students (e.g., Neal & Schanzenbach, 2010), and (b) including students who score some distance around proficiency on the prior state test (e.g., Ladd & Lauen, 2010; Springer, 2008b).

To compare GCPS' prescriptive labels with quantiles from prior year, I divided the distribution of students' prior year state test scores into deciles and then calculated the percent of each decile that was labeled each of MYP, Priority, and Enrichment (shown in Figure 18 for the Benchmark C labels). The lowest-performing students from the prior year are nearly completely identified as low-performing by the district, but the middle and high-performers show more volatility with benchmark performance. For example, in both subjects, more than 85% of the students whose prior scores place them in the lowest two deciles received the MYP label, indicating that students who scored low on the prior year's state test continue to score low by district standards. A relatively sizeable proportion of students from the middle deciles (i.e., bubble students on the prior state test) are over-identified, meaning their benchmark performance is lower than indicated by the prior year. For instance, in math, 55% of students in the 4th decile, 38% in the 5th decile, and 27% in the 6th decile started the year as bubble students but were labeled by the district in February as low-performers (with similar percentages in reading). The district-assigned labels indicate that high performance does not necessarily carry over the following year. While 88% (math) and 91% (reading) of students who score in the top decile are identified as Enrichment, the eighth and ninth deciles show that many students who were high performing in the prior year drop in status by the following year. Performance in math appears more unstable than in reading. In math, only 55% of students who score in the ninth decile and 29% of students in the eighth decile are assigned the Enrichment label. In reading, those percentages are 71% and 42%, respectively. These decile graphs indicate that student performance varies between the prior year state test and progress towards the current year's standards.

In the second comparison, I applied the same definition as Ladd and colleagues to identify grade-level (bubble) students as those who score between -0.5 and 0.5 standard deviations (SD) around the proficiency line, with students scoring -0.5 SD and below considered Low performers, and students scoring 0.5 SD and above considered High performers (Ladd & Lauen, 2010; Lauen & Gaddis, 2012). The bar graphs in Figure 19 show what percent of students would be labeled low, grade-level, and high based on the prior year were assigned the MYP, Priority, and Enrichment label by the district and have similar patterns to those in the decile graphs. Of those students prior research would have identified as grade-level (bubble) by being within 0.5 SD of proficiency, only about 2/3 of them were assigned the Priority label in both subjects. About 20% of students considered grade level the previous year dropped and are labeled by the district as MYP, with 10% (math) and 14% (reading) of those students labeled Enrichment. About ¾ of the students who scored more than 0.5 SD below proficiency in the prior year were labeled MYP, with nearly all of the remaining students assigned the Priority label. Seventy-six percent (reading) and 70% (math) of students who scored more than 0.5 SD above proficiency were labeled Enrichment, with nearly all of the remaining high performing students receiving the Priority label.

These comparisons indicate that student performance is variable over time. Using these strategies on prior year test scores to identify bubble, low, and high students would lead to some misidentification regarding which students schools would consider bubble students, at least in GCPS. Absent from these comparisons, however, are the large percentage of students who did not have a prior year test score, which would eliminate them from analyses relying on that metric to identify bubble students. This is highlighted in Figure 20, which shows similar bar graphs to those just described except they now include what percentage of students labeled MYP, Priority, and Enrichment were missing that metric.

Figure 20 shows that a large proportion of the students who were labeled Priority would be misidentified by using prior year test scores to define them. Only 56% (math) and 60% (reading) would have been considered bubble students based on the distance from proficiency from the previous year. That means nearly four in ten of the students whom GCPS considered bubble students—including about 18% who were missing prior year test score—would be misidentified or left out of these analyses. Between 16% and 34% of students labeled as MYP and Enrichment are missing their prior year test score.

Taken together, these comparisons of the district's labels with the methods commonly used by prior researchers indicate that there is variation in student performance over the course of the school year. A sizeable proportion of bubble students from the prior year, perhaps unsurprising given that status, move up and down in the distribution and end up being labeled as low- or high-performers. One of the biggest issues this comparison highlights is the large proportion of students who would simply be left out of analyses using prior test scores because they do not have that data. The majority of those students with missing data are those taking the state test for the first year in elementary school (in this case, third-graders). While there is little researchers can do to work around missing data like this in administrative datasets, these students represent a substantial number of the actual students that GCPS labeled as Priority. Especially in 2012-13, when the state accountability system prioritized third- and seventh-graders by rating schools on these grades separately, not having data to include third-grade students in analyses would limit the ability to accurately identify the effect of being a bubble student.

Prior research into triage may not have found evidence of triage using administrative data because the bubble students were not accurately identified in the data. The variation in performance shown between the state test in one year and student benchmark scores in the next points to using more local assessment information when possible to more plausibly determine whom schools are likely to view as bubble students.

Another lesson for researchers is that elementary and middle schools in GCPS behaved differently. The quantitative results indicated that middle schools were more responsive to accountability incentives, and TAIP middle schools were more likely than elementary schools to use TAIP funds to facilitate triage. Disaggregating analyses by school level may reveal differences not present when data are combined.

Policymakers

The implications for policymakers have been highlighted in the discussions for each chapter. Broadly, this dissertation adds to the literature which indicates that incentives matter. Student outcomes varied based on the metrics used to rate schools and educators. Using proficiency rates to assess schools resulted in a focus on students close to proficiency. Changing the incentives led to changed behavior and better outcomes for lower-performing students. The policy changes included (a) removing required consequences for failing to meet state-determined proficiency rates, (b) including value-add metrics to principal and teacher evaluation systems, and (c) having schools overseen by the district rather than the state. An additional incentive change which singled out individual grades led to schools providing students in those grades with special treatment while neglecting students in other grades. State policymakers had equity intentions, and this work points to the importance of aligning the metrics used to rate schools and educators with the intentions.

District Leaders

The results from these analyses have implications for district leaders who share data with schools, and the conversations the GCPS research team had in response to these findings help illustrate those implications. The results across the analysis chapters indicate that the labels assigned to students on their benchmarks matter. It is clear from the TAIP documents that schools used the district-supplied test-score labels to identify students who would receive special treatment, meaning the labels matter because schools used them. It is unclear, however, if the students themselves responded to the labels in ways that affected their state test scores. As discussed in Chapter 2, the research team talked about no longer assigning test score labels to students based on their benchmarks because it appeared those labels affected student outcomes. If the labels provide schools with no additional information beyond the raw score and might harm a group of students—especially low performers who need even more additional help—then why even share those labels? These are valuable conversations for district research teams to have about the purpose of providing student data to schools and how to help educators understand what the data do (and do not) say.

Beyond the labels, the messages sent by the district also matter. In regards to TAIP funds, the district encouraged schools to target the interventions to the Priority students. Most schools did so, and the targeted students had better outcomes than the following year, when funds were not available. Yet schools targeted these students in ways that harmed the lowest-performing students. When presented with these results, a member of the GCPS research team responded, "so the good news is that the program worked, and the bad news is that the program worked." That these results are both good news and bad news to the district highlights the complicated nature of this district initiative: they provided schools with a tool to micro-target students based on the accountability system, but that led to schools taking resources away from low- and high-performers.

This reaction highlights another implication for district leaders. The research team originally created the Priority, Multi-Year Plan, and Enrichment labels for internal use to determine which schools should receive the additional TAIP funds. One of the superintendents in GCPS, however, heard about the labels and wanted all schools to focus on the Priority students before the state test. This led to the prescriptive labels being disseminated district-wide and schools being encouraged to offer TAIP interventions to Priority students. It makes sense that the superintendent wanted schools to focus on students closest to the proficiency line because the waiver accountability system held districts accountable primarily based on proficiency rates. But these prescriptive labels and subsequent encouragement to focus on Priority schools sent the message to schools that targeting bubble students was sanctioned by the district[26]. This illustrates district offices which had conflicting agendas in regards to how schools should use benchmark data. Prior research indicates that conflicting agendas at the district level has consequences for instructional improvement (Cobb, Jackson, Henrick, & Smith, 2018). Combined with the lack of guidance regarding how schools should design their TAIP interventions, these conflicting agendas likely contributed to the use of funds to facilitate triage. This highlights the importance of providing guidelines regarding interventions to ensure that students are not harmed by remediation provided to other students. If district leaders want schools to focus on students

[26] Furthermore, the Priority label was clearly viewed as valuable by some schools; the GCPS research director reported that they had to turn off the Target Student Report (Appendix Figure B5) in the data warehouse in the fall of the 2013-14 school year because schools wanted to identify and target Priority students starting at the beginning of the following school year.

across the test score distribution, they need to communicate their vision of remediation with school leaders.

Limitations

This dissertation provides insight into how schools in a large urban district used benchmark data during shifting accountability systems. The benefits of using a single district is the availability of rich contextual information regarding policies and practices. It is limited, however, in its generalizability to other urban districts. While many other districts have adopted benchmark testing systems similar to the one described here (Burch, 2010; Means et al., 2010), GCPS represents a unique case study because they shared prescriptive labels which specifically identified the bubble students and then offered TAIP funding to 29 schools to target Priority students.

Because the waiver accountability system changed multiple incentives statewide simultaneously, it is difficult using this dataset to disentangle which incentives led to improved outcomes for low-performers. In addition, this dissertation compares two old accountability regimes: NCLB and the waivers offered to states in 2011. This state has had two more iterations of accountability policy implemented since the waiver. Furthermore, in 2015, Congress passed and President Obama signed a reauthorization of the Elementary and Secondary Education Act, called the Every Student Succeeds Act (ESSA), which replaces the states' waiver accountability systems. While the accountability regimes have since changed, the new ESSA requires that states use multiple of indicators of academic achievement, and many of them use a combination of proficiency rates and value-add growth metrics.

Further Research

This dissertation offers a look into how schools responded to benchmark data across accountability systems and leads to multiple avenues for further research. The quantitative analyses were limited to looking for effects on student test scores, but the qualitative work provides a view into the effect of policies and programs on students that go beyond test scores. More work of this nature would be useful to provide a better sense of how these policies play out on the ground. Student attendance and discipline data are often available in quantitative datasets and could offer a broader perspective on the student experience. In addition, Chapter 2 found different outcomes for elementary and middle school students. The findings suggested that middle schools were more responsive to accountability incentives, but this is worth further study. For example, middle schools appeared to have focused on Basic students over Below Basic students during NCLB. Is that because middle school teachers hold different views on these students than do elementary schools, or did both sets of teachers try to focus on these students, but it is easier in middle than elementary schools? Examining differences in teachers' attitudes and disaggregating quantitative analyses across school levels could help answer these questions.

This project found more consistent and stronger quantitative results in math than in reading, but the TAIP documents indicated that schools did not distinguish their interventions between the subjects. Were test scores different by subject because it is easier to "teach to the test" in math than in reading? Prior research indicates that state tests assess a smaller percentage of the math standards than reading standards (Jennings & Bearak, 2014), which may make it easier to behave strategically in math than in reading. This is worth further research.

The TAIP documents revealed that schools designed targeted interventions in a variety of ways. While these analyses could not quantify the effects of these different interventions, the interventions were unlikely to be equally effective. Given the press to effectively remediate for students who are struggling, this project points to further research into how to structure interventions that are targeted to student need.

This project was unique within the triage literature because it uses three different benchmark scores that schools had access to at different points in the school year. While the quantitative analyses simply substituted the new benchmark score for the old one, an open question is how schools use multiple pieces of data throughout the year. How do teachers and school leaders integrate new testing data with prior student information? In addition, it is worth exploring the cumulative effect of being labeled low- or high-performing on student outcomes.

Thinking about the cumulative effect of these benchmark labels also leads to questions regarding the effect of the benchmark data on students. Do students know their labels? Are they aware of their relative achievement level compared to their classmates? How does being told that they are “Below Basic” or “Advanced” influence students’ views of self and their motivation? Test scores are used to calculate the effect of these labels on these outcomes, but the practical effects on students are worth consideration. Interviews or focus groups with students would broaden the conversation to better understand how students are affected by being placed in interventions like these (or by being excluded from them).

This work indicates that schools responded to incentive changes that occurred when the waiver was adopted. As already noted, this work could not disentangle the effects of multiple policy changes that occurred simultaneously. Because other states tinkered with their accountability systems, both during the waiver and with the new ESSA, further research could separate the effects of different policy changes to identify which ones create perverse incentives and which benefit low-performers.

Conclusion

Accountability has evolved since No Child Left Behind was implemented in 2003. The new ESSA system continues the trend of incentive shifts by adding additional indicators to school accountability ratings that may help paint a more complete picture of school quality than simply student test score results. When ESSA was signed into law in 2015, nearly one hundred researchers, educators, and policymakers signed an open letter to the Department of Education asking that the policy not require states to use proficiency rates to measure school performance because of the perverse incentives created by those metrics (Polikoff, 2016). If state policymakers want students to grow every year, do proficiency rates make sense for measuring school performance? Given these perverse incentives—and this dissertation's contribution to the research which shows evidence that schools did in fact engage in triage—do the benefits of using proficiency rates outweigh the downside that this results in schools neglecting low and high students to focus on those close to proficiency?

A member of the GCPS research team said they believed triage was being reduced in their schools in part because the accountability system for schools is so complicated that "they don't know how to game the system anymore" (GCPS research team meeting). If a murky accountability system leads to more equitable behavior because a clear accountability system leads to a Campbell-like (1959) focus on the measured aspects of the work, policymakers must take this in to account when designing the metrics and incentives of accountability policy. This dissertation shows that schools are responsive to accountability incentives and points to a real opportunity for state policymakers when they design new systems. State policymakers should be aware of the types of perverse incentives in place and work to make the accountability system metrics align with the desired behaviors.

Figures

Figure 18. Comparison of district-assigned prescriptive labels from Benchmark C with deciles of prior year state test

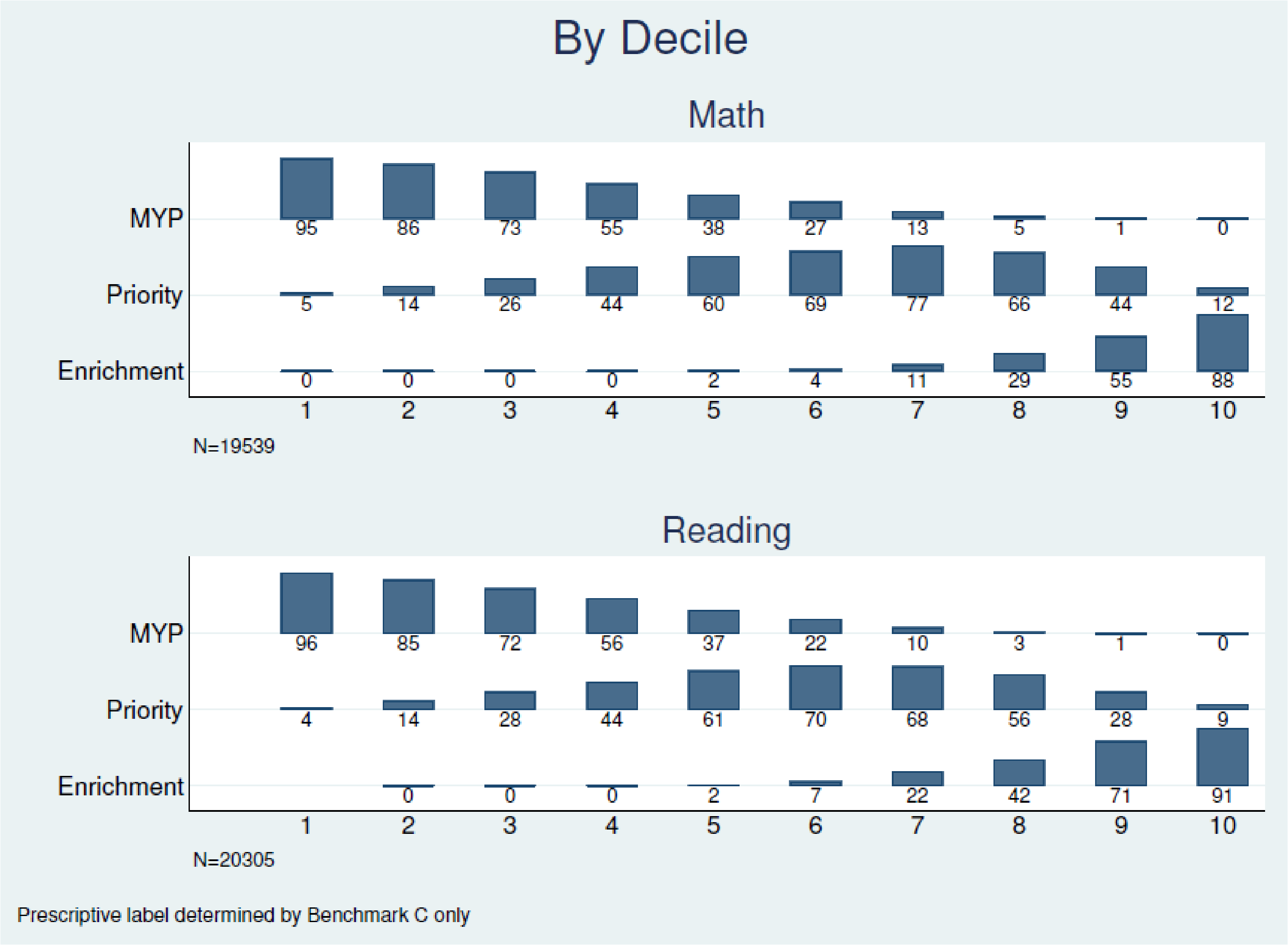


Figure 19. Comparison of district-assigned prescriptive labels with distance from proficiency on prior year state test.

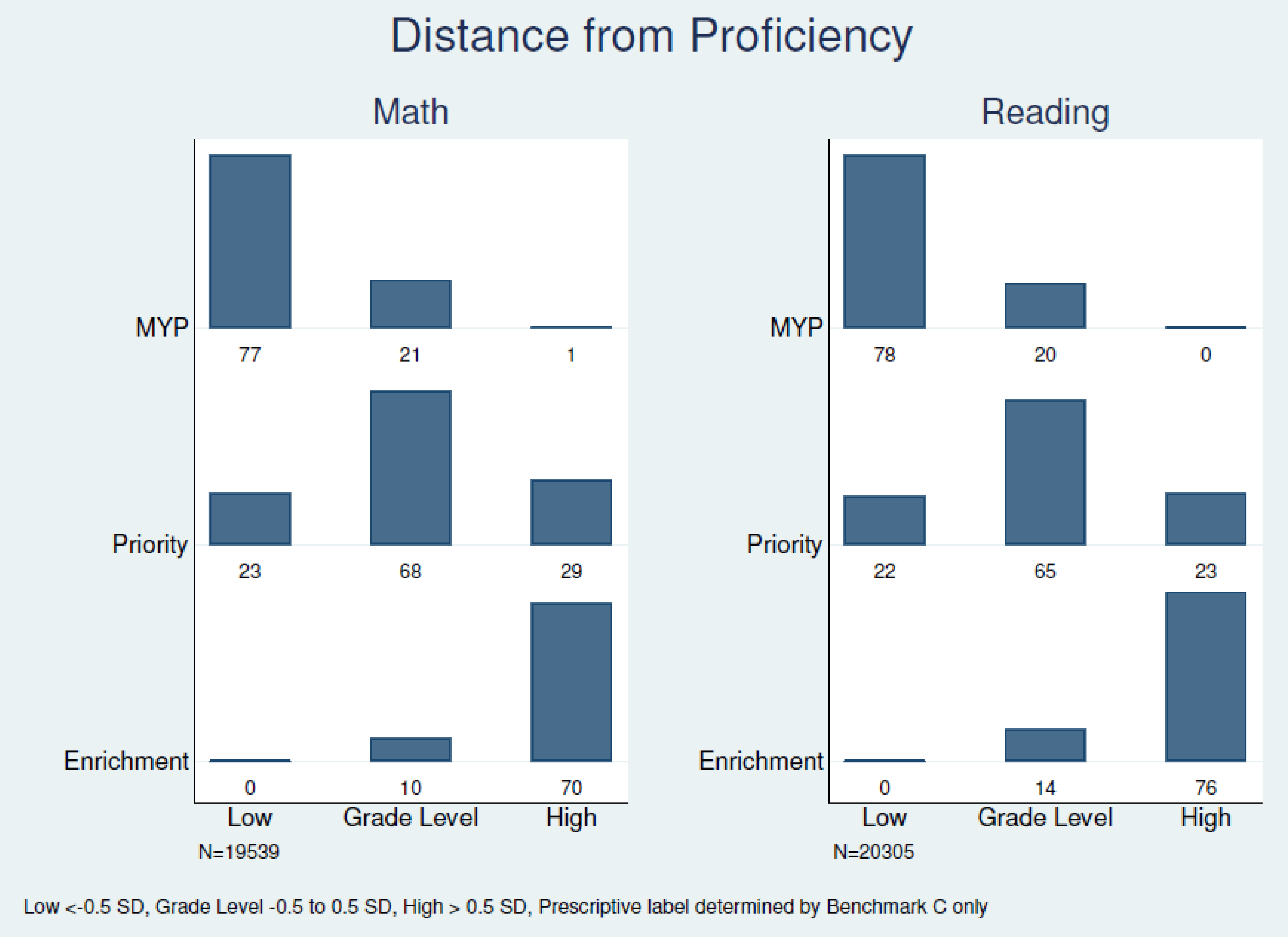
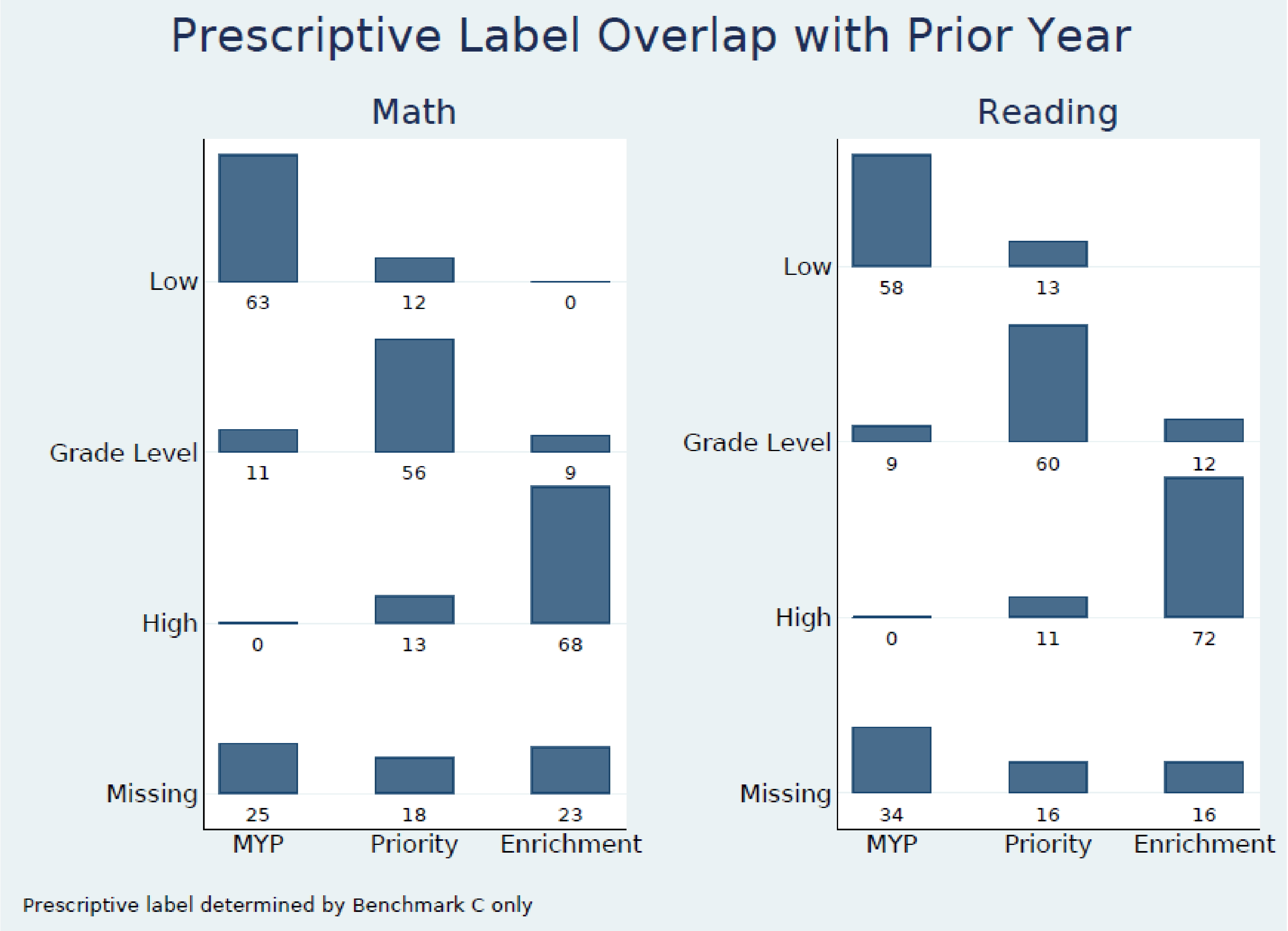


Figure 20. Overlap between district-assigned prescriptive labels and prior year test performance



Appendix A. Supplemental tables and figures from Chapter 2

List of Tables

List of Figures

Appendix Table A1. Estimated discontinuities at label thresholds during NCLB and waiver period, elementary schools math

	Bandwidth (number of questions) →	Below Basic/Basic					Basic/Proficient					Proficient/Advanced				
		2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
Benchmark A	T	-0.018	-0.016	-0.006	-0.008	-0.007	-0.022	-0.002	0.006	0.008	0.016	-0.085	-0.029	0.007	0.002	0.004
		(0.034)	(0.024)	(0.021)	(0.019)	(0.017)	(0.028)	(0.022)	(0.019)	(0.017)	(0.016)	(0.046)	(0.036)	(0.029)	(0.027)	(0.025)
	T x waiver	-0.013	-0.011	-0.018	-0.019	-0.020	-0.022	-0.002	-0.013	-0.015	-0.016	0.057	-0.004	0.001	-0.002	0.005
		(0.027)	(0.023)	(0.021)	(0.021)	(0.021)	(0.027)	(0.024)	(0.021)	(0.020)	(0.020)	(0.045)	(0.038)	(0.035)	(0.032)	(0.030)
	Obs	8645	11627	13686	15080	15372	8153	11203	13780	15706	16805	3380	4832	6477	7867	8854
	R^2	0.453	0.471	0.488	0.506	0.511	0.458	0.491	0.523	0.547	0.563	0.419	0.455	0.486	0.511	0.532
Benchmark B	T	-0.043	0.022	0.017	0.033	0.040*	-0.054	-0.027	-0.025	-0.009	-0.000	0.033	-0.008	0.005	-0.018	-0.025
		(0.035)	(0.027)	(0.023)	(0.021)	(0.019)	(0.033)	(0.024)	(0.021)	(0.019)	(0.018)	(0.046)	(0.034)	(0.029)	(0.026)	(0.024)
	T x waiver	-0.048	-0.040	-0.041	-0.046	-0.045	-0.007	-0.022	-0.013	-0.025	-0.029	0.027	0.055	0.042	0.046	0.037
		(0.032)	(0.030)	(0.027)	(0.025)	(0.025)	(0.026)	(0.023)	(0.021)	(0.021)	(0.021)	(0.037)	(0.031)	(0.028)	(0.028)	(0.027)
	Obs	6595	8883	11049	12610	13946	7317	10243	12979	15148	16317	3963	5573	7156	8689	9861
	R^2	0.406	0.427	0.454	0.472	0.493	0.427	0.464	0.503	0.536	0.549	0.436	0.476	0.505	0.536	0.552
Benchmark C	T	-0.010	0.004	0.007	0.023	0.033	-0.084**	-0.048*	-0.045*	-0.041*	-0.037*	0.020	-0.002	0.000	0.010	0.020
		(0.040)	(0.031)	(0.027)	(0.025)	(0.024)	(0.031)	(0.022)	(0.019)	(0.018)	(0.017)	(0.040)	(0.029)	(0.025)	(0.022)	(0.021)
	T x waiver	-0.048	-0.052	-0.066*	-0.071**	-0.068**	-0.012	-0.023	-0.032	-0.037	-0.025	0.005	0.025	0.031	0.034	0.032
		(0.036)	(0.031)	(0.028)	(0.027)	(0.025)	(0.026)	(0.023)	(0.022)	(0.021)	(0.020)	(0.031)	(0.028)	(0.027)	(0.025)	(0.025)
	Obs	5916	8050	10083	11839	12993	7169	9814	12384	14590	15919	4540	6265	7968	9555	10753
	R^2	0.394	0.426	0.467	0.496	0.508	0.418	0.456	0.504	0.545	0.565	0.451	0.486	0.525	0.554	0.577

Notes: *p<0.05; **p<0.01, ***p<0.001. "T" represents the discontinuity in outcomes for students who received the higher of the two labels. Standard errors in parentheses are clustered at the school by grade by year level. All models condition on student covariates (race, ethnicity, FRPL, ELL, and disability status), school covariates (log enrollment, prior year's percent proficient in math, percent of school that is black, Hispanic, white, FRPL, ELL, and has a disability), grade by year fixed effects, and school fixed effects.

Appendix Table A2. Estimated discontinuities at label thresholds during NCLB and waiver period, middle schools math

	Bandwidth (number of questions) →	Below Basic/Basic					Basic/Proficient					Proficient/Advanced				
		2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
Benchmark A	T	-0.011	-0.002	-0.002	-0.002	0.008	0.004	0.019	0.014	0.026*	0.028*	0.050	0.019	0.024	0.020	0.019
		(0.023)	(0.017)	(0.015)	(0.014)	(0.013)	(0.023)	(0.017)	(0.014)	(0.013)	(0.012)	(0.029)	(0.021)	(0.018)	(0.017)	(0.016)
	T x waiver	0.023	0.016	0.001	0.002	0.002	-0.005	-0.017	-0.033*	-0.037**	-0.036*	0.022	0.001	-0.018	-0.023	-0.026
		(0.019)	(0.017)	(0.016)	(0.016)	(0.015)	(0.017)	(0.015)	(0.014)	(0.014)	(0.014)	(0.024)	(0.021)	(0.020)	(0.019)	(0.019)
	Obs	22409	30078	35580	38849	40883	18829	26184	33241	37674	39724	8867	12490	16332	20065	22079
	R^2	0.402	0.426	0.450	0.469	0.482	0.453	0.487	0.521	0.537	0.542	0.394	0.438	0.477	0.515	0.533
Benchmark B	T	-0.014	-0.001	0.006	0.017	0.018	-0.010	0.024	0.036**	0.039***	0.036**	0.059*	0.028	0.029	0.021	0.025
		(0.024)	(0.018)	(0.015)	(0.014)	(0.013)	(0.020)	(0.015)	(0.013)	(0.012)	(0.011)	(0.026)	(0.019)	(0.017)	(0.016)	(0.015)
	T x waiver	-0.026	-0.032*	-0.032*	-0.032*	-0.031*	-0.013	-0.031*	-0.044***	-0.038**	-0.036**	0.036	0.021	0.024	0.011	0.006
		(0.018)	(0.016)	(0.015)	(0.015)	(0.015)	(0.015)	(0.014)	(0.013)	(0.013)	(0.013)	(0.020)	(0.018)	(0.016)	(0.016)	(0.016)
	Obs	18740	25621	31482	35405	37423	20159	27630	33934	38279	40518	11818	16597	21288	24799	26273
	R^2	0.394	0.425	0.458	0.484	0.501	0.467	0.498	0.526	0.541	0.547	0.460	0.496	0.527	0.552	0.568
Benchmark C	T	0.049*	0.052**	0.056***	0.056***	0.049***	0.018	0.002	-0.002	0.006	0.013	0.023	-0.005	0.000	0.008	0.012
		(0.022)	(0.017)	(0.013)	(0.013)	(0.012)	(0.018)	(0.014)	(0.012)	(0.011)	(0.010)	(0.025)	(0.018)	(0.016)	(0.014)	(0.014)
	T x waiver	-0.038*	-0.059***	-0.058***	-0.064***	-0.067***	-0.029	-0.024	-0.029*	-0.030*	-0.032*	-0.022	-0.000	0.014	0.015	0.011
		(0.018)	(0.016)	(0.015)	(0.015)	(0.015)	(0.016)	(0.014)	(0.013)	(0.012)	(0.013)	(0.018)	(0.017)	(0.015)	(0.015)	(0.015)
	Obs	16540	22588	28244	32310	34647	17838	24686	31124	35989	38794	12668	17422	22034	26061	27766
	R^2	0.404	0.436	0.473	0.491	0.502	0.423	0.463	0.507	0.536	0.554	0.461	0.505	0.540	0.572	0.589

Notes: *p<0.05; **p<0.01, ***p<0.001. "T" represents the discontinuity in outcomes for students who received the higher of the two labels. Standard errors in parentheses are clustered at the school by grade by year level. All models condition on student covariates (race, ethnicity, FRPL, ELL, and disability status), school covariates (log enrollment, prior year's percent proficient in math, percent of school that is black, Hispanic, white, FRPL, ELL, and has a disability), grade by year fixed effects, and school fixed effects.

Appendix Table A3. Estimated discontinuities at label thresholds during NCLB and waiver period, elementary schools reading

		Below Basic/Basic					Basic/Proficient					Proficient/Advanced				
Bandwidth (number of questions) →		2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
Benchmark A	T	0.019	0.026	0.015	0.020	0.031	-0.006	-0.002	0.010	0.012	0.015	0.056	0.032	0.010	0.010	0.013
		(0.043)	(0.031)	(0.027)	(0.024)	(0.024)	(0.033)	(0.023)	(0.019)	(0.018)	(0.016)	(0.042)	(0.031)	(0.025)	(0.023)	(0.022)
	T x waiver	0.014	-0.003	0.007	-0.008	-0.018	0.030	0.013	0.016	0.022	0.017	-0.034	-0.024	-0.008	0.002	0.001
		(0.037)	(0.033)	(0.030)	(0.028)	(0.026)	(0.024)	(0.021)	(0.020)	(0.018)	(0.018)	(0.034)	(0.030)	(0.028)	(0.027)	(0.026)
	Obs	4470	6238	7819	9417	10914	6436	8804	11158	13466	15199	4331	5859	7251	8527	9763
	R^2	0.402	0.417	0.436	0.459	0.480	0.397	0.444	0.488	0.534	0.569	0.402	0.448	0.488	0.516	0.540
Benchmark B	T	0.022	-0.009	-0.004	0.008	0.026	-0.020	-0.016	-0.023	-0.020	-0.021	-0.043	0.017	0.024	0.030	0.029
		(0.047)	(0.033)	(0.027)	(0.024)	(0.023)	(0.031)	(0.023)	(0.020)	(0.017)	(0.016)	(0.036)	(0.026)	(0.023)	(0.020)	(0.019)
	T x waiver	-0.034	-0.019	-0.024	-0.040	-0.042	-0.001	-0.005	0.002	-0.003	-0.005	-0.071*	-0.070**	-0.048	-0.057*	-0.059*
		(0.034)	(0.030)	(0.026)	(0.024)	(0.023)	(0.026)	(0.022)	(0.020)	(0.019)	(0.018)	(0.032)	(0.027)	(0.025)	(0.025)	(0.025)
	Obs	4794	6595	8304	9897	11142	6705	9322	11855	13961	15693	4675	6492	8136	9607	10734
	R^2	0.369	0.408	0.437	0.474	0.491	0.411	0.453	0.505	0.549	0.582	0.440	0.473	0.505	0.540	0.557
Benchmark C	T	0.023	0.003	0.013	-0.001	0.013	0.025	0.030	0.021	0.017	0.008	0.019	-0.004	0.030	0.035	0.037
		(0.050)	(0.034)	(0.029)	(0.026)	(0.024)	(0.032)	(0.023)	(0.020)	(0.018)	(0.016)	(0.038)	(0.026)	(0.023)	(0.022)	(0.020)
	T x waiver	-0.023	-0.040	-0.040	-0.022	-0.028	0.001	-0.001	0.003	0.003	0.007	-0.021	-0.017	-0.023	-0.031	-0.024
		(0.037)	(0.032)	(0.029)	(0.026)	(0.026)	(0.025)	(0.023)	(0.020)	(0.019)	(0.019)	(0.031)	(0.029)	(0.029)	(0.028)	(0.027)
	Obs	3796	5347	6823	8303	9670	6679	9272	11767	14067	15420	4584	6122	7605	8987	10212
	R^2	0.426	0.434	0.451	0.474	0.495	0.414	0.450	0.506	0.558	0.584	0.455	0.489	0.516	0.542	0.560

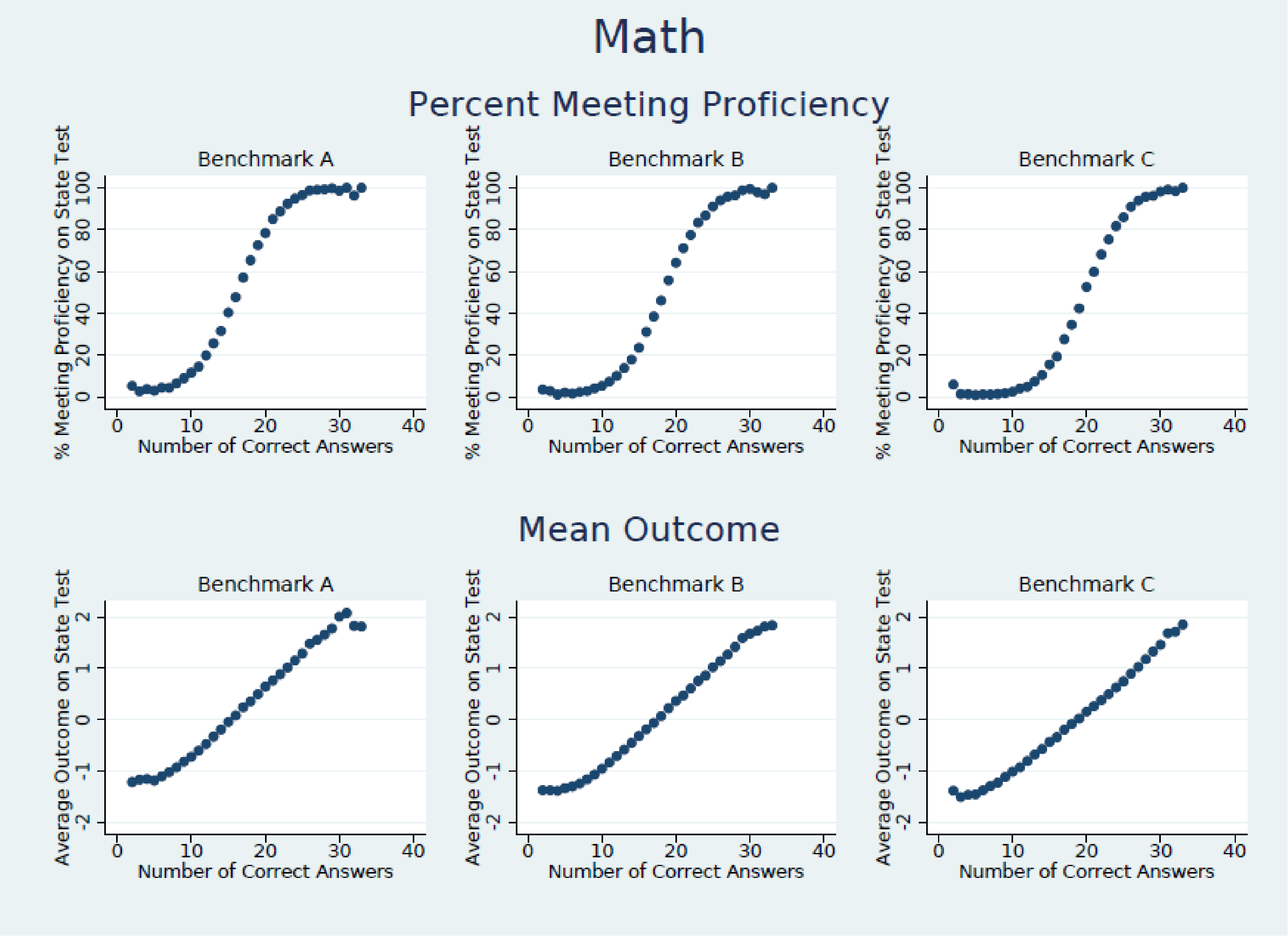
Notes: *p<0.05; **p<0.01, ***p<0.001. “T” represents the discontinuity in outcomes for students who received the higher of the two labels. Standard errors in parentheses are clustered at the school by grade by year level. All models condition on student covariates (race, ethnicity, FRPL, ELL, and disability status), school covariates (log enrollment, prior year’s percent proficient in reading, percent of school that is black, Hispanic, white, FRPL, ELL, and has a disability), grade by year fixed effects, and school fixed effects.

Appendix Table A3. Estimated discontinuities at label thresholds during NCLB and waiver period, middle schools reading

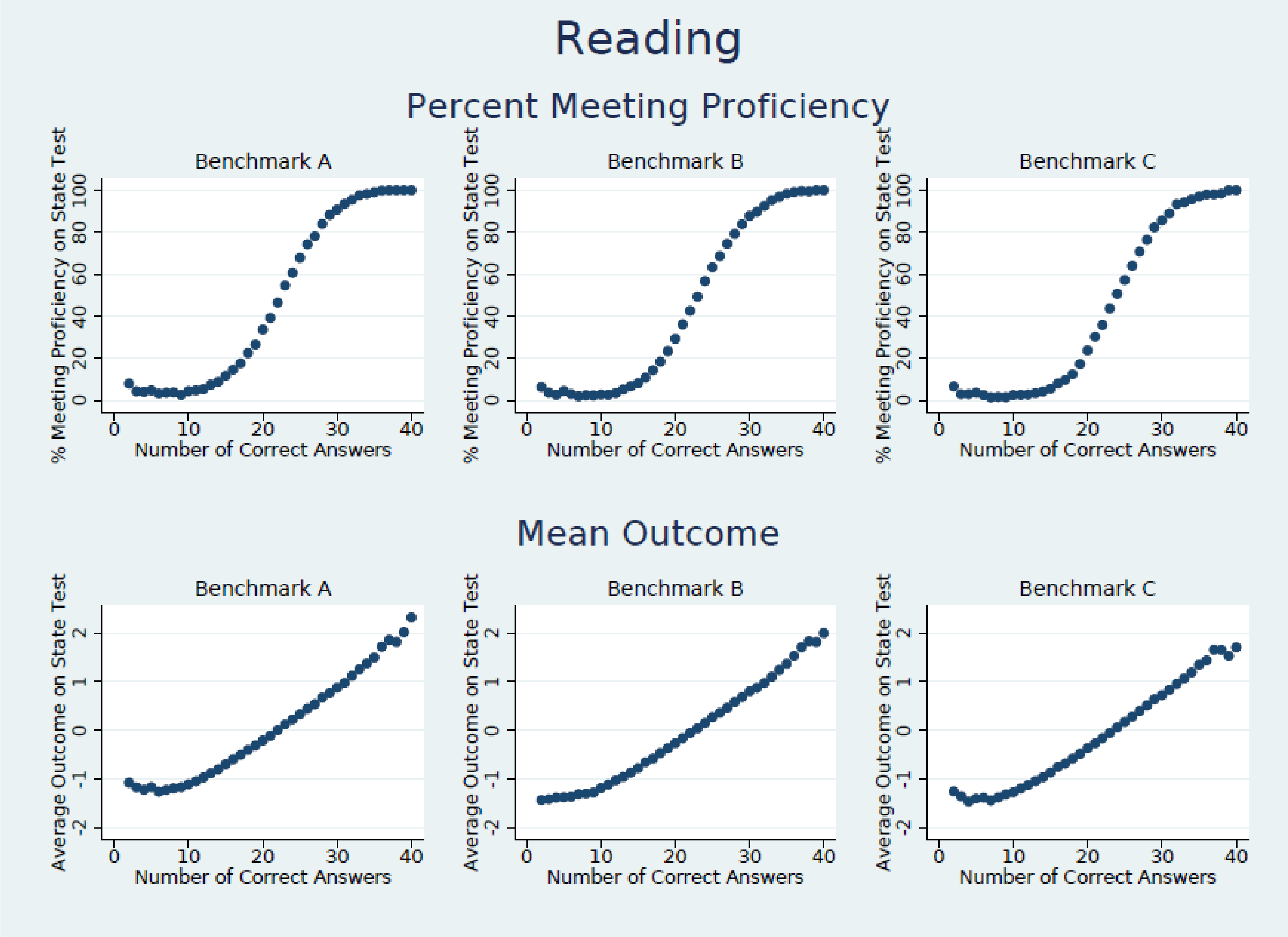
		Below Basic/Basic					Basic/Proficient					Proficient/Advanced				
	Bandwidth (number of questions) →	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
Benchmark A	T	0.005	-0.019	-0.007	0.003	0.009	-0.003	0.005	0.011	0.014	0.016	0.004	-0.003	0.002	0.012	0.009
		(0.032)	(0.022)	(0.019)	(0.017)	(0.016)	(0.020)	(0.015)	(0.012)	(0.011)	(0.010)	(0.025)	(0.018)	(0.015)	(0.014)	(0.013)
	T x waiver	-0.027	-0.023	-0.013	-0.024	-0.029*	-0.021	-0.021	-0.015	-0.014	-0.014	-0.001	0.003	-0.005	-0.001	0.002
		(0.021)	(0.017)	(0.016)	(0.015)	(0.014)	(0.014)	(0.012)	(0.011)	(0.011)	(0.010)	(0.022)	(0.019)	(0.017)	(0.017)	(0.017)
	Obs	11059	15362	19461	23331	27118	17170	23724	29763	35527	40333	10663	14790	18507	22017	25382
	R^2	0.390	0.414	0.431	0.454	0.480	0.448	0.481	0.522	0.559	0.587	0.427	0.472	0.508	0.535	0.557
Benchmark B	T	-0.006	0.005	0.006	0.005	0.019	-0.014	-0.023	-0.013	-0.016	-0.017	0.001	-0.006	0.002	0.016	0.015
		(0.028)	(0.019)	(0.017)	(0.015)	(0.014)	(0.019)	(0.014)	(0.012)	(0.011)	(0.010)	(0.025)	(0.018)	(0.016)	(0.014)	(0.014)
	T x waiver	0.015	0.007	-0.000	-0.016	-0.019	0.022	0.017	0.011	0.012	0.013	0.016	-0.002	-0.004	-0.015	-0.020
		(0.021)	(0.018)	(0.016)	(0.015)	(0.014)	(0.015)	(0.013)	(0.012)	(0.011)	(0.011)	(0.020)	(0.019)	(0.018)	(0.017)	(0.018)
	Obs	10881	15094	19140	22700	26057	17492	24178	30548	36285	40686	11284	15501	19347	23112	26578
	R^2	0.378	0.405	0.429	0.451	0.470	0.447	0.491	0.527	0.564	0.595	0.447	0.485	0.514	0.543	0.566
Benchmark C	T	0.035	0.027	0.020	0.035*	0.034*	-0.033	-0.027*	-0.021	-0.012	-0.011	-0.007	-0.005	-0.000	0.001	0.008
		(0.030)	(0.021)	(0.018)	(0.016)	(0.015)	(0.018)	(0.013)	(0.011)	(0.010)	(0.009)	(0.024)	(0.018)	(0.015)	(0.014)	(0.013)
	T x waiver	-0.057*	-0.042*	-0.047*	-0.050**	-0.060**	-0.010	-0.017	-0.017	-0.023*	-0.026**	0.045*	0.055**	0.048**	0.038*	0.033*
		(0.024)	(0.021)	(0.019)	(0.018)	(0.018)	(0.014)	(0.012)	(0.011)	(0.010)	(0.010)	(0.020)	(0.019)	(0.018)	(0.017)	(0.017)
	Obs	9633	13496	17302	21149	24413	18276	25153	31690	37475	42381	10901	14944	18629	22259	25544
	R^2	0.436	0.455	0.477	0.503	0.514	0.454	0.494	0.540	0.577	0.611	0.448	0.486	0.521	0.551	0.569

Notes: *p<0.05; **p<0.01, ***p<0.001. “T” represents the discontinuity in outcomes for students who received the higher of the two labels. Standard errors in parentheses are clustered at the school by grade by year level. All models condition on student covariates (race, ethnicity, FRPL, ELL, and disability status), school covariates (log enrollment, prior year’s percent proficient in reading, percent of school that is black, Hispanic, white, FRPL, ELL, and has a disability), grade by year fixed effects, and school fixed effects.

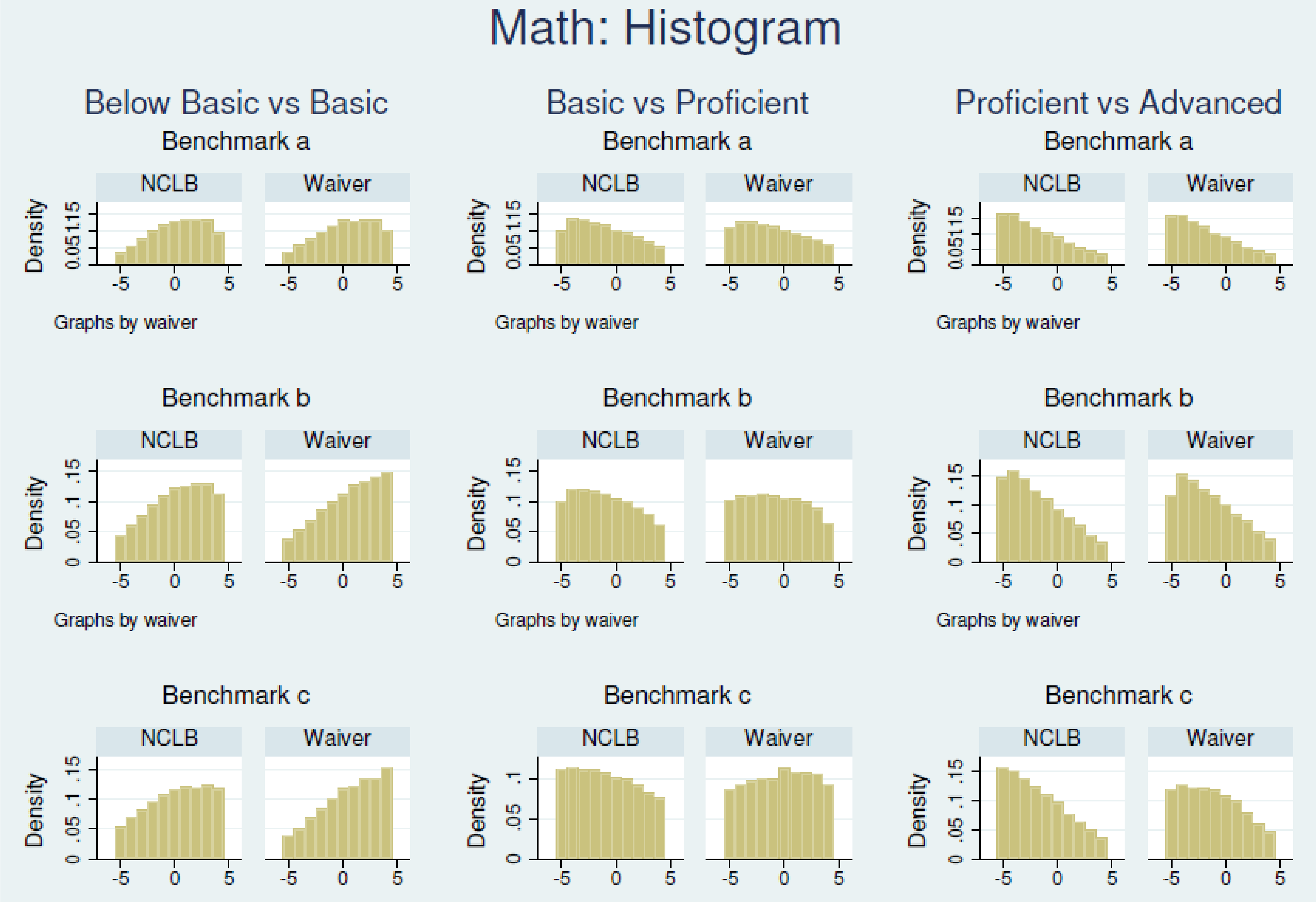
Appendix Figure A1. Percent of students meeting proficiency and average outcome by benchmark rawscore, math



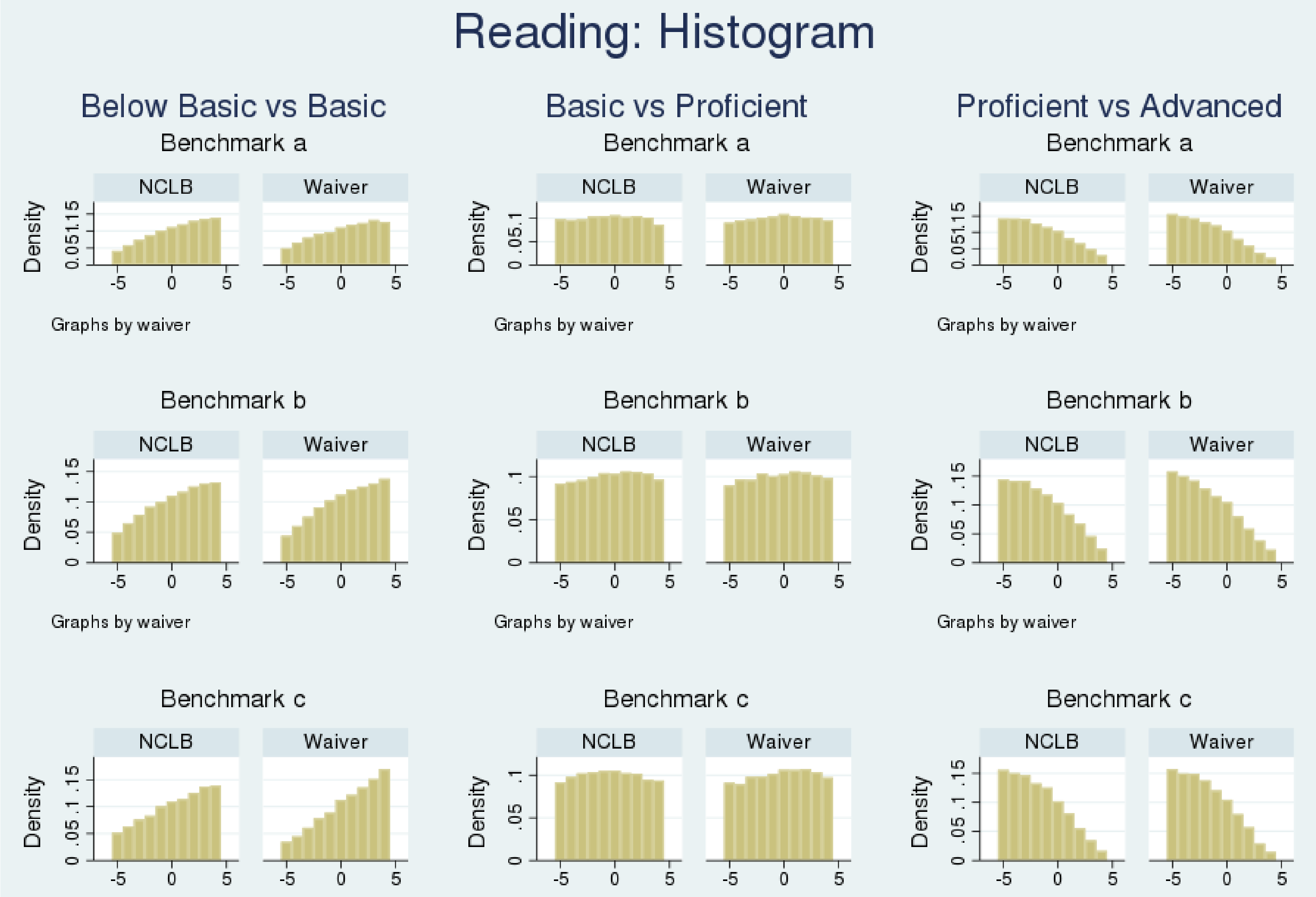
Appendix Figure A2. Percent of students meeting proficiency and average outcome by benchmark rawscore, reading



Appendix Figure A3. Histogram of number of questions answered correctly during NCLB and waiver, math



Appendix Figure A4. Histogram of number of questions answered correctly during NCLB and waiver, reading



Appendix B. Supplemental tables and figures from Chapter 3

List of Tables

List of Figures

Appendix Table B1. Effect of Chapter 2 labels when using difference-in-differences models

	Math		Read	
	Benchmark B	Benchmark C	Benchmark B	Benchmark C
Below Basic	-0.061***	-0.043***	-0.069***	-0.083***
	(0.008)	(0.008)	(0.007)	(0.008)
Below Basic x Waiver	0.071***	0.047***	0.037***	0.011
	(0.010)	(0.012)	(0.010)	(0.013)
Proficient	0.016*	-0.014	-0.017**	-0.005
	(0.008)	(0.007)	(0.006)	(0.006)
Proficient x Waiver	-0.044***	-0.030***	0.003	-0.014*
	(0.009)	(0.009)	(0.007)	(0.007)
Advanced	0.001	0.004	0.046***	0.070***
	(0.015)	(0.014)	(0.011)	(0.013)
Advanced x Waiver	-0.020	-0.039**	-0.007	0.006
	(0.015)	(0.014)	(0.014)	(0.014)
Observations	122518	123496	125501	125564
Adjusted R^2	0.75	0.75	0.77	0.77

Notes: *p<0.05; **p<0.01, ***p<0.001. Standard errors in parentheses are clustered at the school by grade level. All models condition on student testing variables up to cubic polynomials and interactions between all the variables (prior year state test z-score; projected percentile; raw scores for Benchmarks A, B, and C, centered on proficiency), student demographics (race, ethnicity, FRPL, ELL, and disability status), school covariates (log enrollment, prior year's percent proficient in that subject, percent of school that is black, Hispanic, white, FRPL, ELL, and has a disability), school fixed effects, and grade by year fixed effects.

Appendix Table B2. Consistency of Multi-Year Plan, Priority, and Enrichment labels across Benchmarks B and C

Panel A.

		Math Benchmark C				
		MYP	Priority	Enrichment	No Label	Missing
Math Benchmark B	MYP	8,063	2,544	99	36	1,139
	Priority	1,184	6,205	1,400	58	1,083
	Enrichment	21	474	3,294	40	287
	No Label	10	100	112	57	12
	Missing	1,034	1,300	527	7	1,779

Panel B. Reading

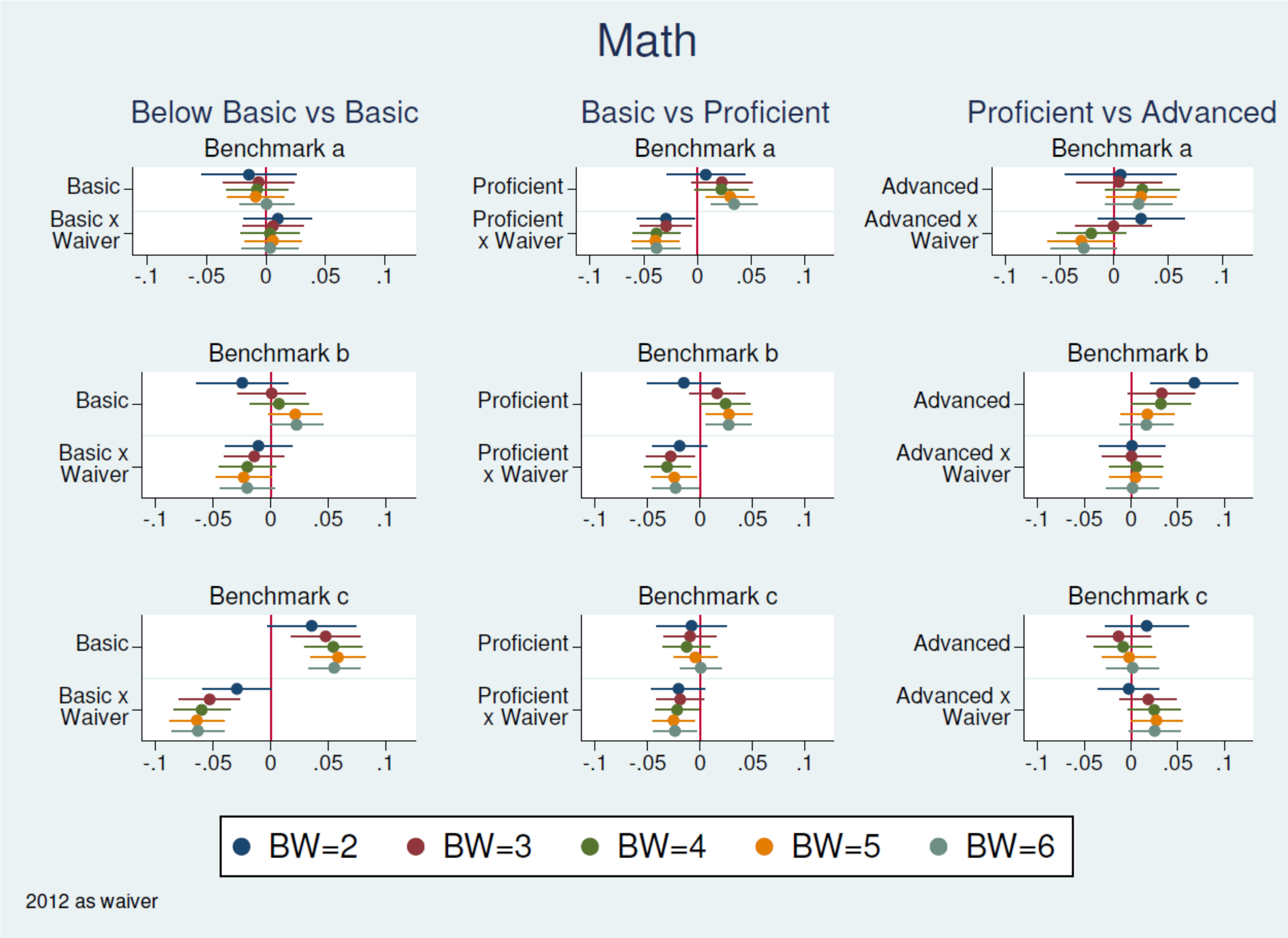
		Reading Benchmark C				
		MYP	Priority	Enrichment	No Label	Missing
Reading Benchmark B	MYP	8,789	2,251	67	4	1,186
	Priority	1,466	5,330	1,326	21	1,130
	Enrichment	24	922	4,083	17	511
	No Label	1	43	36	7	6
	Missing	1,028	1,295	602	9	1,778

Appendix Table B3. Estimated discontinuities at Basic/Proficient threshold by TAIP funding and 2012-13, reading

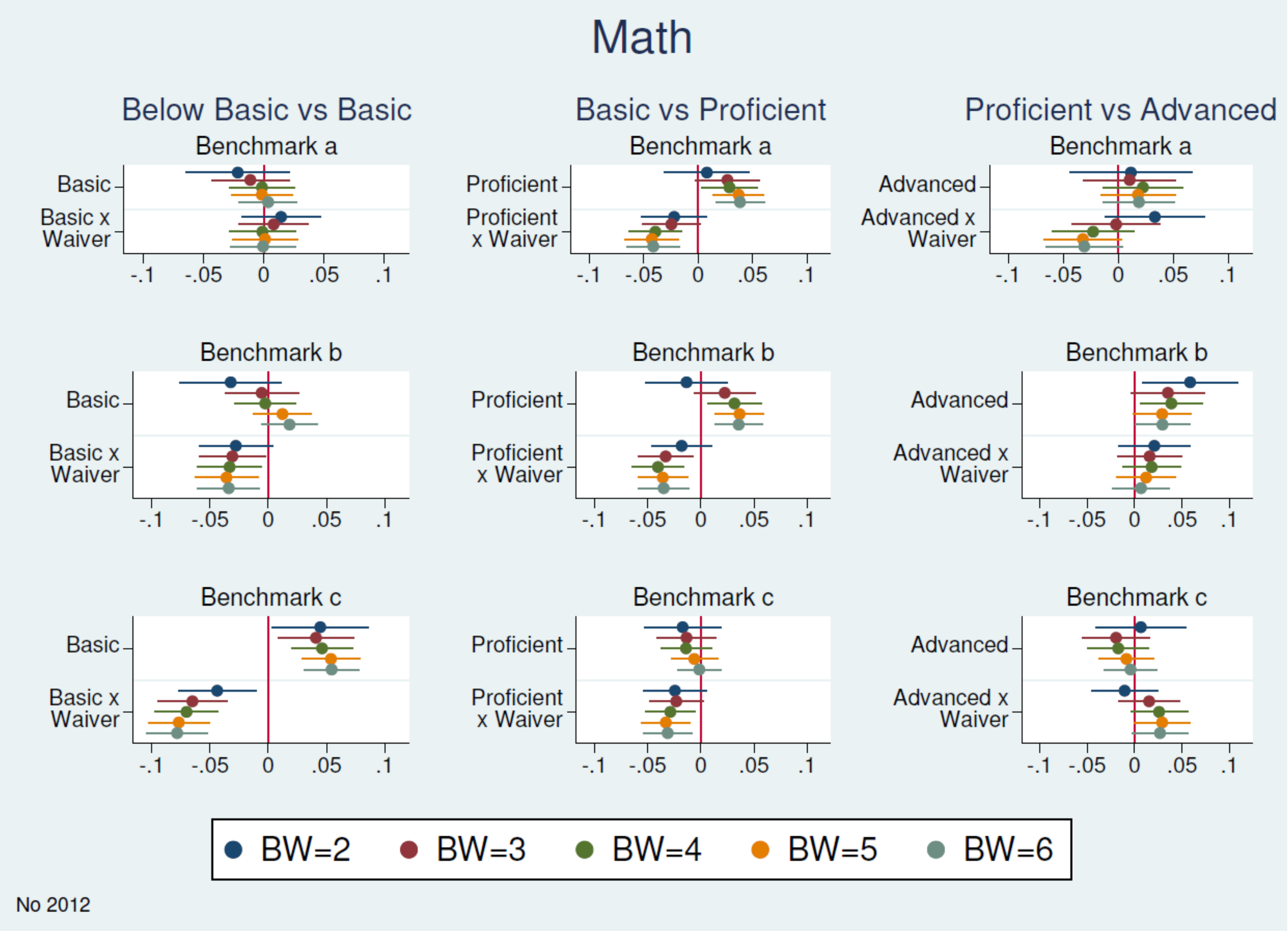
	Benchmark B					Benchmark C				
Bandwidth (number of questions) →	2	3	4	5	6	2	3	4	5	6
Proficient	-0.006	-0.014	-0.012	-0.016	-0.015	-0.018	-0.012	-0.005	0.002	0.001
	(0.018)	(0.014)	(0.012)	(0.011)	(0.010)	(0.017)	(0.013)	(0.011)	(0.010)	(0.010)
Proficient x has TAIP	-0.019	-0.016	-0.007	-0.003	-0.006	0.002	-0.001	-0.012	-0.014	-0.013
	(0.015)	(0.013)	(0.011)	(0.011)	(0.010)	(0.015)	(0.013)	(0.012)	(0.011)	(0.011)
Proficient x waiver	0.016	0.011	0.017	0.017	0.010	0.013	-0.004	-0.014	-0.016	-0.015
	(0.024)	(0.020)	(0.018)	(0.017)	(0.017)	(0.022)	(0.018)	(0.017)	(0.015)	(0.014)
Has TAIP x waiver	-0.007	0.003	0.034	0.035	0.026	-0.008	-0.018	-0.011	0.007	0.001
	(0.030)	(0.026)	(0.023)	(0.022)	(0.021)	(0.029)	(0.026)	(0.023)	(0.021)	(0.020)
Proficient x has TAIP x waiver	0.005	-0.010	-0.035	-0.030	-0.023	-0.025	-0.014	-0.008	-0.014	-0.016
	(0.037)	(0.030)	(0.027)	(0.027)	(0.025)	(0.032)	(0.028)	(0.024)	(0.022)	(0.021)
Proficient x 2013	-0.034	-0.023	-0.027	-0.022	-0.012	-0.033	-0.011	0.003	0.004	0.002
	(0.029)	(0.023)	(0.021)	(0.020)	(0.019)	(0.026)	(0.023)	(0.021)	(0.018)	(0.017)
Has TAIP x 2013	-0.042	-0.039	-0.059*	-0.051*	-0.044	0.007	0.024	0.006	-0.002	-0.003
	(0.035)	(0.030)	(0.027)	(0.025)	(0.024)	(0.035)	(0.030)	(0.027)	(0.026)	(0.024)
Proficient x has TAIP x 2013	0.060	0.067	0.085*	0.064*	0.056	0.031	0.011	0.018	0.013	0.013
	(0.045)	(0.037)	(0.034)	(0.032)	(0.030)	(0.040)	(0.035)	(0.031)	(0.028)	(0.027)
Observations	24197	33500	42403	50246	56379	24955	34425	43457	51542	57801
R^2	0.436	0.479	0.520	0.559	0.591	0.447	0.486	0.535	0.576	0.605

Notes: *p<0.05; **p<0.01, ***p<0.001. Standard errors in parentheses are clustered at the school by grade by year level. All models condition on student covariates (race, ethnicity, FRPL, ELL, and disability status), school covariates (log enrollment, prior year's percent proficient in reading, percent of school that is black, Hispanic, white, FRPL, ELL, and has a disability), grade by year fixed effects, and school fixed effects.

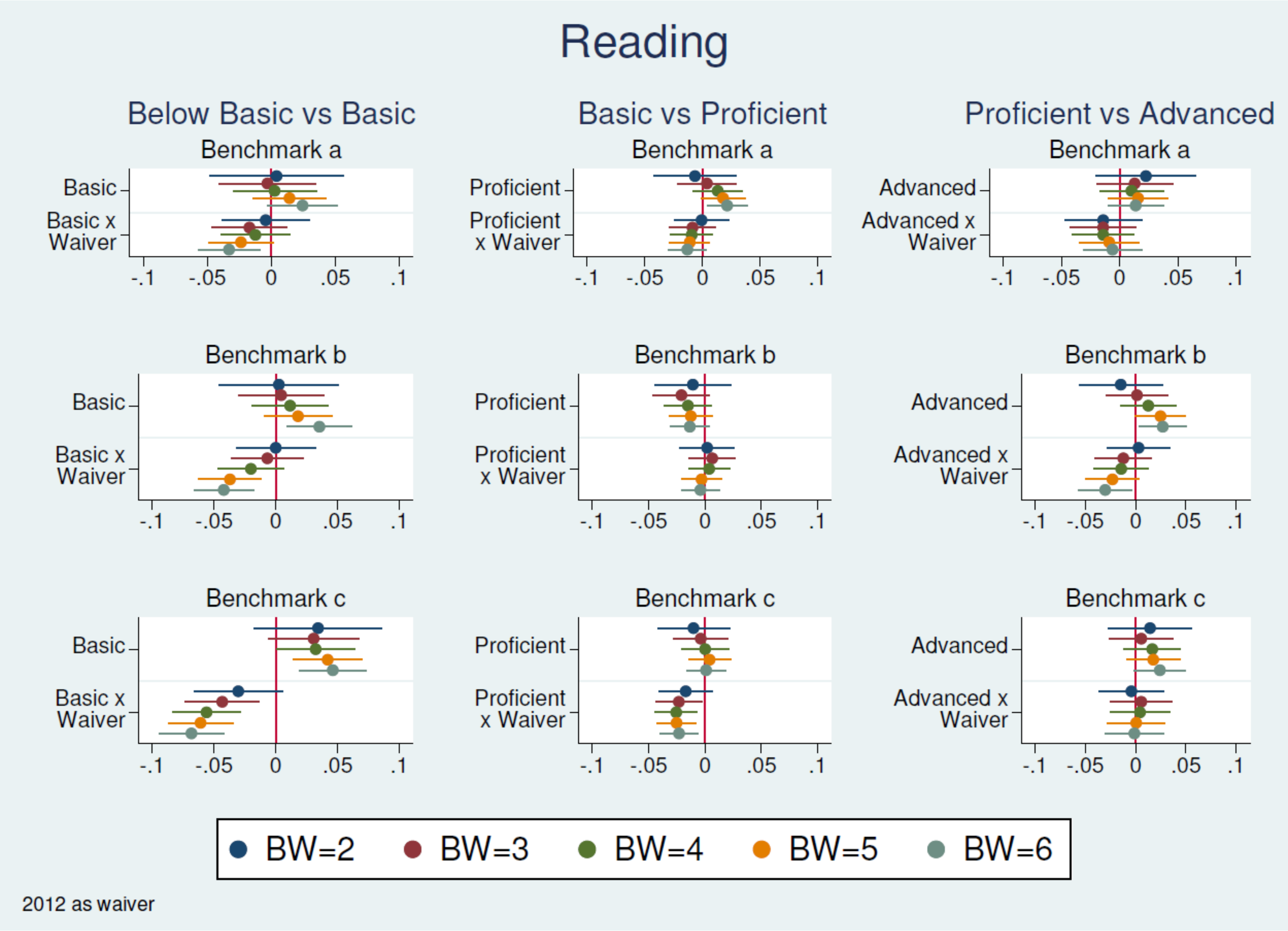
Appendix Figure B1. Estimated discontinuities at label thresholds for math, 2011-12 coded as waiver



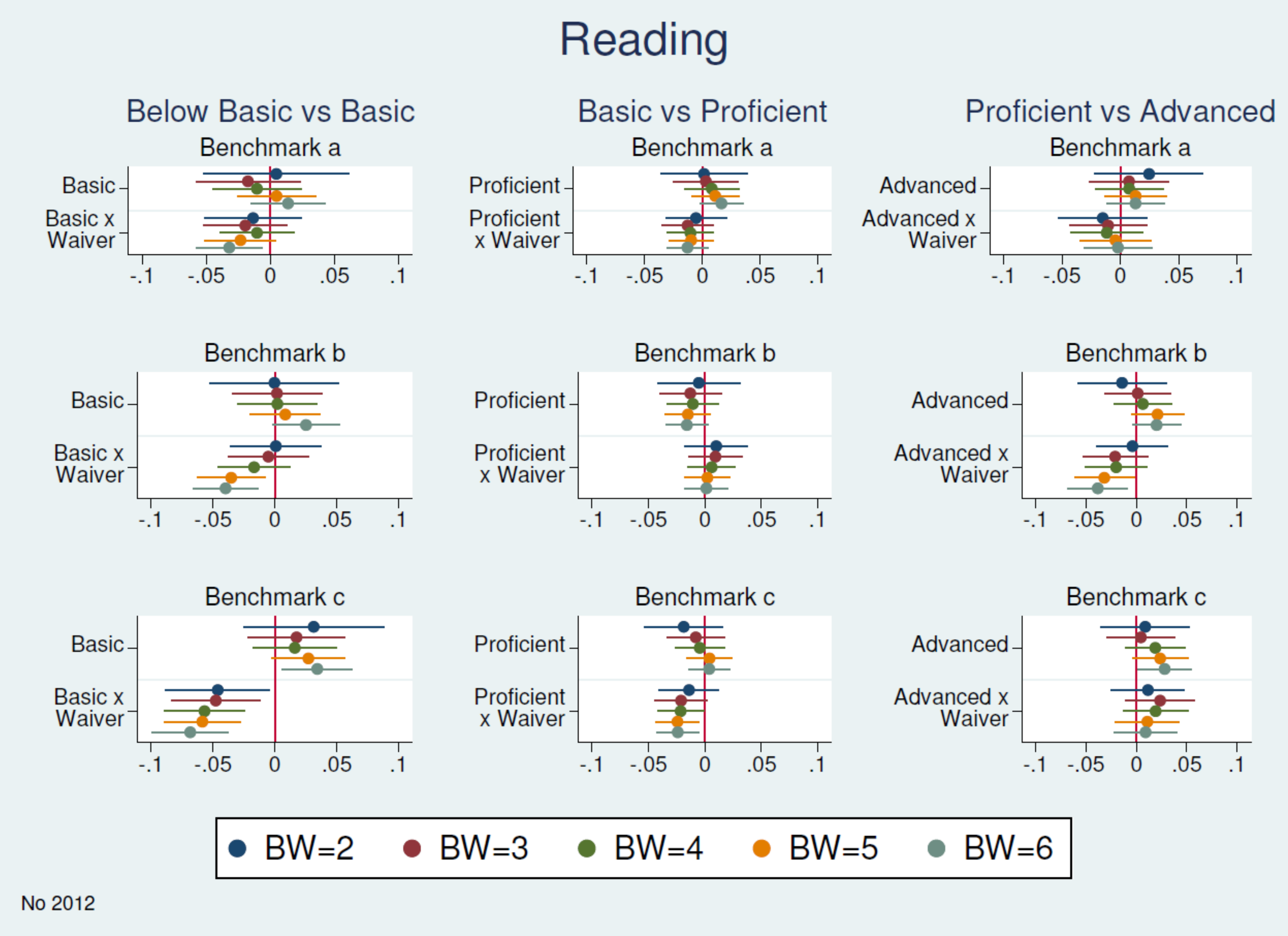
Appendix Figure B2. Estimated discontinuities at label thresholds for math, 2011-12 dropped from analysis



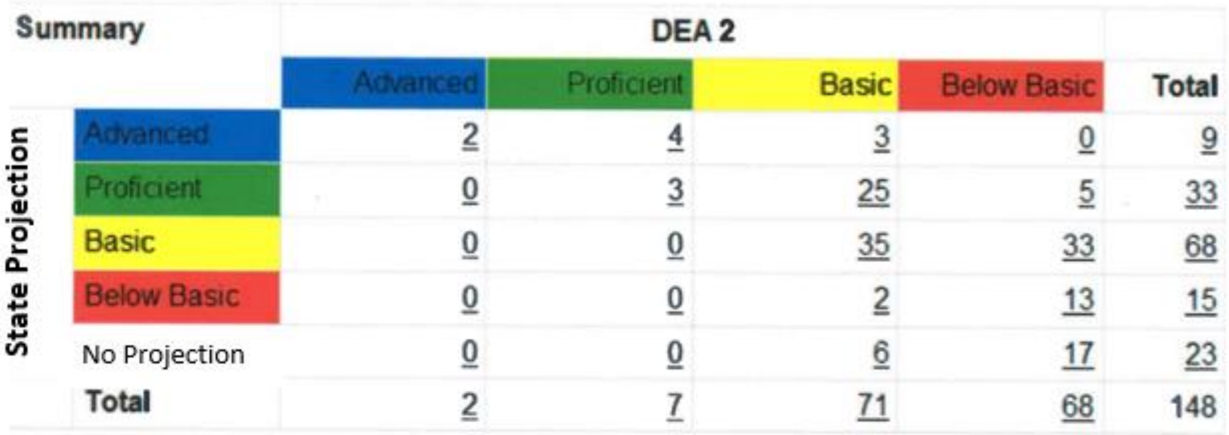
Appendix Figure B3. Estimated discontinuities at label thresholds for reading, 2011-12 coded as waiver



Appendix Figure B4. Estimated discontinuities at label thresholds for reading, 2011-12 dropped from analysis

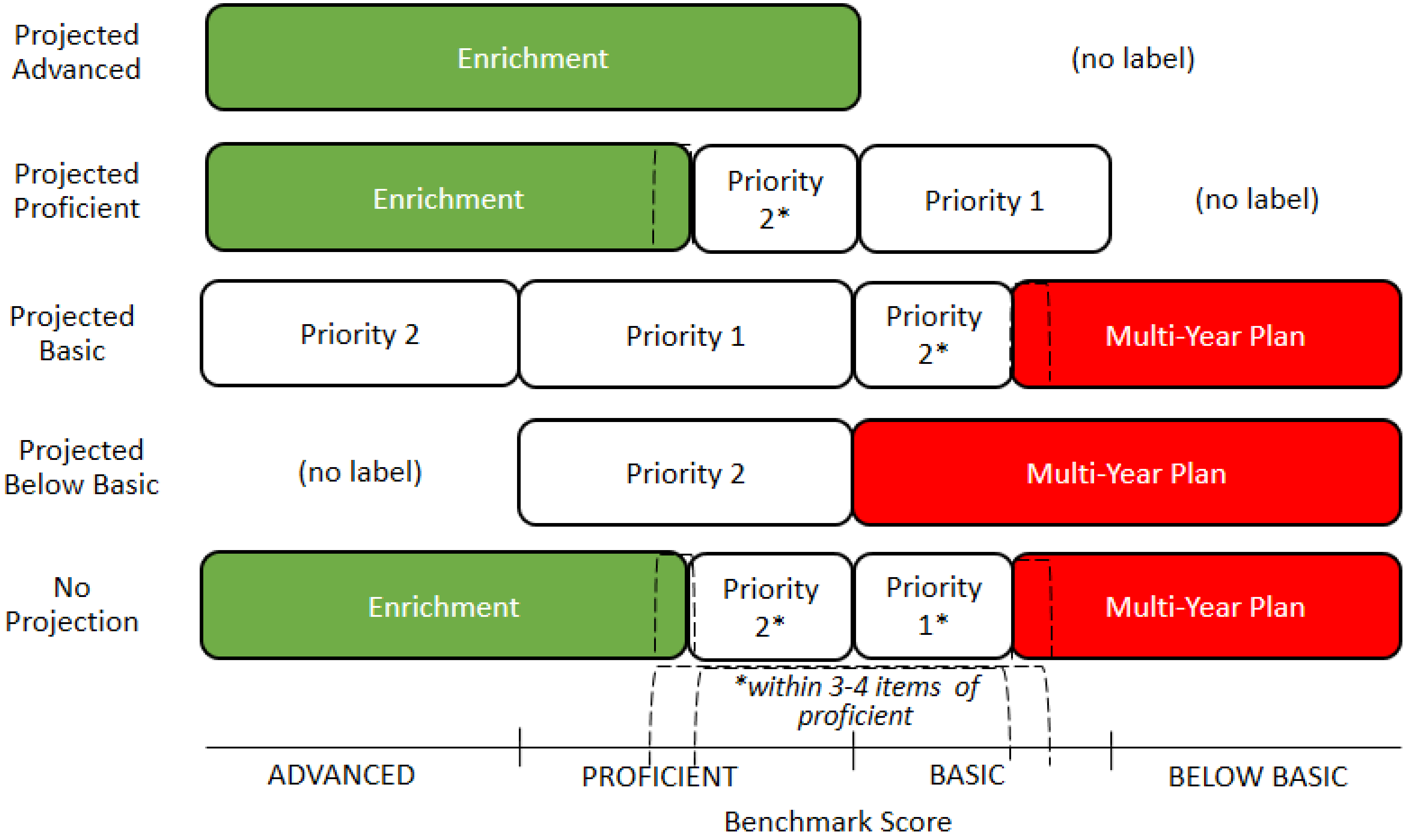


Appendix Figure B5. Virtual data wall report used with *Identification of Target Students* document



Summary		DEA 2				
		Advanced	Proficient	Basic	Below Basic	Total
State Projection	Advanced	2	4	3	0	9
	Proficient	0	3	25	5	33
	Basic	0	0	35	33	68
	Below Basic	0	0	2	13	15
	No Projection	0	0	6	17	23
	Total	2	7	71	68	148

Appendix Figure B6. Stylized version of *Identification of Target Students* matrix



Note: Original *Identification of Target Student* matrix shown in Figure 2.

REFERENCES

Balfanz, R., Legters, N., West, T. C., & Weber, L. M. (2007). Are NCLB's measures, incentives, and improvement strategies the right ones for the nation's low-performing high schools? *American Educational Research Journal, 44*(3), 559–593. https://doi.org/10.3102/0002831207306768

Ballou, D., & Springer, M. G. (2016). Has NCLB encouraged educational triage? Accountability and the distribution of achievement gains. *Education Finance and Policy*. https://doi.org/10.1162/EDFP_a_00189

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal*, *42*(2), 231–268. https://doi.org/10.3102/00028312042002231

Brown, A. B., & Clift, J. W. (2010). The unequal effect of adequate yearly progress: Evidence from school visits. *American Educational Research Journal*, *47*(4), 774–798. https://doi.org/10.3102/0002831210374644

Bulkley, K. E., Christman, J. B., Goertz, M. E., & Lawrence, N. R. (2010). Building with benchmarks: The role of the district in Philadelphia's benchmark assessment system. *Peabody Journal of Education*, *85*(2), 186–204. https://doi.org/10.1080/01619561003685346

Burch, P. (2010). The bigger picture: Institutional perspectives on interim assessment technologies. *Peabody Journal of Education*, *85*(2), 147–162. https://doi.org/10.1080/01619561003685288

Burns, M. K., & Gibbons, K. (2012). *Implementing Response-to-Intervention in Elementary and Secondary Schools*. New York: Routledge.

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, *2*(1), 67–90. https://doi.org/10.1016/0149-7189(79)90048-X

Cobb, P., Jackson, K., Henrick, E. C., & Smith, T. (2018). *Systems for instructional improvement: Creating coherence from the classroom to the district office*. Cambridge, MA: Harvard Education Press.

Cullen, J. B., & Reback, R. (2006). Tinkering toward accolades: School gaming under a performance accountability system (No. w12286). *National Bureau of Economic Research*. https://doi.org/10.1016/S0278-0984(06)14001-8

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

Davidson, K. L., & Frohbieter, G. (2011). *District adoption and implementation of interim and benchmark assessments. (CRESST Report 806)*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Dee, T., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, *30*(3), 418–446. https://doi.org/10.1002/pam

Dee, T., Jacob, B., & Schwartz, N. L. (2012). The effects of NCLB on school resources and practices. *Educational Evaluation and Policy Analysis*, *35*(2), 252–279. https://doi.org/10.3102/0162373712467080

Diamond, J. B., & Spillane, J. P. (2004). High-stakes accountability in urban elementary schools. *Teachers College Record*, *106*(6), 1145–1176. Retrieved from http://www3.interscience.wiley.com/journal/118771481/abstract

Discovery Education Assessment. (n.d.). *Discovery Education Assessment Research*. Retrieved from http://www.discoveryeducation.com/pdf/assessment/Discovery_Education_Assessment_Research.pdf

Elementary and Secondary Education Act [ESEA] Flexibility Request. (2012). Retrieved from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html

Elementary and Secondary Education Act [ESEA] Flexibility Request. (2015). Retrieved from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html

Figlio, D. N. (2006). Testing, crime and punishment. *Journal of Public Economics*, *90*(4–5), 837–851. https://doi.org/10.1016/j.jpubeco.2005.01.003

Figlio, D. N., & Getzler, L. S. (2006). Accountability, ability and disability: Gaming the system? In T. J. Gronberg & D. W. Jansen (Eds.), *Improving School Accountability (Advances in Applied Microeconomics, Volume 14)*. Bingley, UK: Emerald Group Publishing Limited.

Gillborn, D., & Youdell, D. (1999). *Rationing education: Policy, practice, reform, and equity*. United Kingdom: McGraw-Hill Education.

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

Horn, I. S. (2016). Accountability as a design for teacher learning: Educators' sensemaking about mathematics and equity in the NCLB era. *Urban Education*. https://doi.org/10.1177/0042085916646625

Horn, I. S., Kane, B. D., & Wilson, J. (2015). Making sense of student performance data: Data use logics and mathematics teachers' learning opportunities. *American Educational Research Journal*, *52*(2), 208–242. https://doi.org/10.3102/0002831215573773

Imbens, G. W., & Kalyanaraman, K. (2009). Optimal bandwidth choice for the regression discontinuity estimator (No. 14726). *National Bureau of Economic Research*. https://doi.org/10.1017/CBO9781107415324.004

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*(2), 615–635. https://doi.org/10.1016/j.jeconom.2007.05.001

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

Jennings, J. L., & Bearak, J. M. (2014). "Teaching to the Test" in the NCLB era: How test predictability affects our understanding of student performance. *Educational Researcher*, *43*(8), 381–389. https://doi.org/10.3102/0013189X14554449

Jennings, J. L., & Sohn, H. (2014). Measure for measure: How proficiency-based accountability systems affect inequality in academic achievement. *Sociology of Education*, *87*(2), 125–141. https://doi.org/10.1177/0038040714525787

Krieg, J. (2008). Are students left behind? The distributional effects of the No Child Left Behind Act. *Education Finance and Policy*, *3*(2), 250–281. https://doi.org/10.1162/edfp.2008.3.2.250

Ladd, H. F., & Lauen, D. L. (2010). Status versus growth : The distributional effects of school accountability policies. *Journal of Policy Analysis and Management*, *29*(3), 426–450. https://doi.org/10.1002/pam

Lauen, D. L., & Gaddis, S. M. (2012). Shining a light or fumbling in the dark? The effects of NCLB's subgroup-specific accountability on student achievement. *Educational Evaluation* *and Policy Analysis*, *34*(2), 185–208. https://doi.org/10.3102/0162373711429989

Manna, P. (2011). *Collision Course: Federal Education Policy Meets State and Local Realities*. Washington, D.C.: CQ Press.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, *142*(2), 698–714. https://doi.org/10.1016/j.jeconom.2007.05.005

McGuinn, P. J. (2006). *No Child Left Behind and the Transformation of Federal Education Policy, 1965-2005*. Lawrence, KS: University Press of Kansas.

McMurrer, J. (2007). Choice, changes, and challenges: Curriculum and instruction in the NCLB era. *Center on Education Policy*.

Means, B., Padilla, C., Gallagher, L., & SRI International. (2010). *Use of education data at the local level: From accountability to instructional improvement*. *U.S. Department of Education, Office of Planning, Evaluation, and Policy Development*. Washington, D.C.

Miller, R. T., Murnane, R. J., & Willett, J. B. (2008). Do Teacher Absences Impact Student Achievement? Longitudinal Evidence From One Urban School District. *Educational Evaluation and Policy Analysis*, *30*(2), 181–200. https://doi.org/10.3102/0162373708318019

Murnane, R. J., & Papay, J. P. (2010). Teachers' views on No Child Left Behind: Support for the principles, concerns about the practices. *The Journal of Economic Perspectives*, *24*(3), 151–166. Retrieved from http://www.ingentaconnect.com/content/aea/jep/2010/00000024/00000003/art00010

Neal, D., & Schanzenbach, D. (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics*, *92*(May), 263–283.

Nichols, A. (2011). rd 2.0: Revised Stata module for regression discontinuity estimation.

Papay, J. P., Murnane, R. J., & Willett, J. B. (2011). How performance information affects human-capital investment decisions: The impact of test-score labels on educational outcomes (No. w17120). *National Bureau of Economic Research*.

Perlstein, L. (2010). Unintended consequences: High stakes can result in low standards. *American Educator*, *34*(2), 6–9.

Polikoff, M. (2016). *A letter to the U.S. Department of Education*. Retrieved from https://morganpolikoff.com/2016/07/12/a-letter-to-the-u-s-department-of-education

Polikoff, M., McEachin, A., Wrabel, S. L., & Duque, M. (2014). The waive of the future? School accountability in the waiver era. *Educational Researcher*, *43*(1), 45–54. https://doi.org/10.3102/0013189X13517137

Reback, R. (2008). Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics*, *92*(5–6), 1394–1415. https://doi.org/10.1016/j.jpubeco.2007.05.003

Reback, R., Rockoff, J., & Schwartz, H. L. (2011). Under pressure: Job security, resource allocation, and productivity in schools under NCLB (No. w16745). *National Bureau of Economic Research*. Retrieved from http://www.nber.org/papers/w16745

Riddle, W. (2012). Major accountability themes of second-round state applications for NCLB waivers. *Center on Education Policy*. Retrieved from http://files.eric.ed.gov/fulltext/ED531861.pdf

Ryan, J. E. (2010). *Five miles away, a world apart: One city, two schools, and the story of educational opportunity in modern America.* Oxford: Oxford University Press.

Schochet, P., Cook, T. D., Deke, J., Imbens, G. W., Lockwood, J. R., Porter, J., & Smith, J.

(2010). Standards for regression discontinuity designs. *What Works Clearinghouse*. Retrieved from http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=231

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin Company.

Springer, M. G. (2008a). Accountability incentives: Do failing schools practice educational triage (unabridged version). *Education Next*, *8*(1), 74–79.

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

Srikantaiah, D. (2009). How state and federal accountability policies have influenced curriculum and instruction in three states: Common findings from Rhode Island, Illinois, and Washington. *Center on Education Policy*. Retrieved from http://eric.ed.gov/?id=ED513300