Do Standardized Tests Hurt Student Learning? Causal Evidence from a Natural Experiment*

Yan Lau†

April 2017

Abstract

This paper estimates the causal effect of testing on student educational outcomes. Standardized testing has become pervasive in the American schooling system as a means of measuring the performance of students and educators. Parents and other stakeholders often raise concerns that testing has negative psychological and educational consequences for students. Focusing on how testing affects learning, I use participation in the National Assessment of Educational Progress (NAEP) tests as the treatment in a natural experiment. Treatment is randomly assigned because NAEP's sampling procedure randomly selects schools across the United States for testing. Schools not selected into the NAEP sample serve as the control group. To measure educational outcomes, I use school-level results from subsequent state-wide tests in which all schools take part. I find that testing has no statistically meaningful impact on subsequent educational outcomes, suggesting that apprehensions related to students being over-tested are not as worrying as espoused.

Key words: Standardized testing, assessment, test anxiety, NAEP, test scores, education

JEL Codes: I21, I28, H43

*I would like to thank Damon Clark, Janet Currie, Nicholas Lawson, Lars Lefgren, Alex Mas, Harvey Rosen, and Cecilia Rouse for invaluable feedback. Taylor Holdaway and Alicia Toshima provided excellent research assistance. I am grateful to the staff at the US Education Department's Institute of Education Sciences for their help with the data.

†Economics Department, Reed College [yanlau@reed.edu]

1 Introduction

What are the effects of standardized testing on student outcomes? Recent policies such as No Child Left Behind (NCLB) and the Race to the Top grant program specifically emphasize the need for testing as part of education reform. However, there has been a backlash from parents and teachers against testing. As the debate ensues, the question though remains: Do students gain more human capital by taking tests, or does testing actually hurt learning outcomes?

In an era where testing has become the norm, these are important questions with a high degree of policy relevance. Policy makers must design curricula and assessment standards in order to measure the performance of schools, teachers, and students. Test results provide information to teachers and administrators for making resource allocation decisions, such as which subjects or areas to focus on. Teachers must allocate precious class time between testing and other pedagogical activities, and additional testing may crowd out instruction time. District administrators have to placate parents who demand quantitative accountability measures, yet at the same time take issue with the impact of the testing regimen on their children, who suffer from test fatigue and anxiety. These parents, concerned about their children's well-being, are especially weary of standardized tests, and there has been a recent trend of resistance through opting out (Harris, 2015). Politicians on both sides of the aisle complain that students are "over-tested" so much so that the Obama administration announced plans to limit the amount of class time devoted to tests (Zernike, 2015). With so much at stake, estimating the impact of testing on learning is of utmost importance.

This paper estimates the causal effect of testing on educational outcomes. Causal effects are difficult to obtain using survey data because the implementation of testing policy could be endogenous. Randomized control trials investigating the effect of testing have thus far been small in scale, and possibly lack external validity beyond laboratory settings. Instead, to estimate the effect of testing, I make use of a large-scale natural experiment that randomly assigns the treatment of testing to students.

Table 1: The NAEP as a Natural Experiment

Time	Treatment Schools	Control Schools
Late January to Early March	NAEP Testing (Randomly Selected)	No NAEP Testing (Not in NAEP Sample)
April to June (Varies by State)	State-wide Standardized Testing	State-wide Standardized Testing

The National Assessment of Educational Progress (NAEP) is a nationally-administered assessment program used as a common metric to track educational progress and development in the United States. Schools across the country are randomly sampled to participate, and students at selected schools are "treated" with tests in subjects such as math and reading. If not selected into the sample, schools are not tested. The random sampling mechanism of NAEP effectively provides random assignment into the testing treatment group (test-takers in schools randomly chosen to participate as part of the NAEP sample) and the control group (students in schools not selected to participate). To measure outcomes, I use school-level results from subsequent state-wide tests in which all schools take part. Table 1 summarizes the timing of tests for both treatment and control groups.

Using school-level data from 40 states and the District of Columbia, I estimate the effects of testing on math and reading state-wide test scores for grades 4 and 8—the grades that receive NAEP testing—in the 2011 and 2013 rounds of NAEP. I find that testing has no statistically meaningful impact on subsequent state-wide standardized testing results.

The remainder of this paper proceeds as follows. Section 2 introduces the literature on testing and its effects. Section 3 presents background information on the NAEP, the natural experiment I use for identification. I then estimate the effects of testing on education outcomes in Section 4. Section 5 discusses these results further, while Section 6 considers the external validity of my findings. Section 7 concludes.

2 Literature on Testing

To assess the impact of testing, it is instructive to briefly discuss what tests actually are, and how they can be classified in various ways. A test is an assessment device used by educators to measure the human capital stock of students. It is one of many methods available for such measurement purposes—other examples include homework, in-class written exercises, group discussions, and direct questioning or observation by the teacher. A test consists of a set of questions that are usually one of two types: multiple-choice questions or constructed response (open-ended) questions.[1] A greater number of correct responses is associated with higher levels of human capital. Tests may ask the same set of questions to all students taking it (e.g. booklet SATs) or they may draw questions from a larger pool, with students answering only a subset of all possible questions (e.g. the computer adaptive Graduate Record Examinations).

Contents of a test can be "standardized" across varying target groups of students. Classroom-based tests, such as in-class pop quizzes, are administered by individual teachers, and each student in the class receives the same set of questions. School-based tests take place at the school level whereby each student taking that subject at the school receives the same set of questions. Lastly, at an even higher level, population-based tests are tests whose coverage of students can cross school, district, state, and even national boundaries. Examples of such testing programs include the California Standards Tests (CSTs) at the state level, the NAEP at the national level, and the Programme for International Student Assessment at the international level. Most tests colloquially referred to as "standardized tests" fall into this latter category of population-based tests.

Tests are targeted at either a sample of students (e.g. NAEP) or the entire population of students (e.g. CSTs). For tests targeting samples of students constituting only a subset of the entire population, the sample can be selected (often representatively) by test administrators

[1]Some researchers note that the choice of question type (in particular, the use of multiple choice) can have positive and negative consequences for student learning (see Roediger and Marsh, 2005).

(e.g. NAEP), or self-selected by participants themselves (e.g. SATs). Tests targeting the entire student population are usually mandated by fiat (e.g. CSTs).

Education experts note that testing and other evaluation methods serve two main purposes: summative assessment (a.k.a. assessment *of* learning) and formative assessment (a.k.a. assessment *for* learning). Summative assessment refers to the process of measuring human capital with the end-goal of obtaining the measurements. Formative assessment, on the other hand, refers to the added step "when the evidence is actually used to adapt the teaching work to meet the [students' learning] needs" (Black and Wiliam, 1998b). These two roles are not mutually exclusive; in fact, the former is necessary for the latter.

The formative assessment role diagnoses a student's strengths and weaknesses. The communication of this feedback can be carried out under the student's own initiative (e.g. checking solutions of a graded test), but oftentimes, the feedback is transmitted through actions of the teacher. Tests results allow teachers to gauge their students' progress and better understand their learning needs (Black and Wiliam, 1998a). This diagnosis is then used to adjust lesson plans and reallocate class time towards activities that focus on correcting weaknesses. Early literature on formative assessment notes that classroom evaluation practices can affect both student achievement and motivation (Crooks, 1988). As education systems mature, there has been a shift in focus towards developing assessments for formative purposes (James, 2010, p. 164).

On the other hand, many assessments are entirely summative in nature. These tests are mainly used for accountability and accreditation purposes—to evaluate the performance of students, teachers, schools, and even entire countries. As such, they are generally conducted targeting the population level so as to draw complete comparisons. Standardized testing at the population level is almost always purely summative. Results are usually reported in the aggregate (e.g. percentage of proficient students at a given school, the SAT score of an individual student aggregating over questions) and do not allow for formative pedagogy. Moreover, by the time these results are released, students will often have moved on to the

next level of schooling.

Standardized tests can be further classified into tests with stakes and tests with no stakes.[2] Attaching stakes to a test means associating test results with consequential outcomes for the student (e.g. promotion to the next grade) and/or the teacher (e.g. advancing to the next level on the salary scale). Proponents for with-stakes testing argue that they incentivize both students and teachers to increase effort and improve learning outcomes (Santibanez, 2010). Other researchers though have found negative (Jones et al., 1999) or mixed effects (Firestone et al., 1998) resulting from the presence of stakes.

Testing can be beneficial or detrimental to students in terms of numerous outcomes, including human capital accumulation, motivation, and psychological well-being. Positive effects on learning stem from a variety of mechanisms. Students may experience "learning by doing" in that the skills or facts tested are honed and reviewed through the testing process itself. In anticipation of the test, students may increase effort studying, and develop strategies for effective learning in the process. Moreover, the test's coverage serves as a guide for students, helping them understand what the learning expectations are, and how to allocate study time effectively. The corrective feedback provided through formative assessment allows students to reactivate and consolidate knowledge, while at the same time unlearning the incorrect. Test results further serve as signals to students, informing them which topics they excel in, and may help them choose which majors and careers to pursue in the future.

There is an extensive literature on the positive effects of testing, much of which is experimental in nature. Roediger and Karpicke (2006) investigate the "testing effect" by randomly treating one group of participants with tests while another control group was allowed to study the material for the same amount of time. They find that testing improves long-term retention of the material as measured in subsequent follow-up assessments. Carpenter and Kelly (2012) argue that benefits apply not only to simple tasks and recall of facts, but also to higher-order tasks involving the learning of spatial orientation. In an experiment involv-

[2] I use this dichotomy for convenience. The terminology in the literature is not consistent, with references sometimes made to "low-stakes" or "high-stakes."

ing computer-rendered virtual environments, they show that participants who learned the required task through testing perform better compared to participants who learned the task through studying.

Beyond the laboratory, there have also been "field" studies conducted in classroom settings. Gijbels et al. (2005) offered regular written assessments to law school students and find improvements in the final exam scores of those who completed the assessment tasks relative to those who did not. McDaniel et al. (2007) expose a subset of students to weekly quizzes and a control group to additional reading. They find that the test group performed better than the control group in later assessments, and that quizzes containing short answer questions were more beneficial compared to quizzes with multiple-choice questions. In a similar setup, Pennebaker et al. (2013) administer daily computer-based tests for university students, and note that tested students performed better in exams compared to non-tested students. Carpenter et al. (2009) assess the retention of American history knowledge by 8th graders and report that testing improves retention in an assessment conducted nine months later.

On the other hand, tests can have negative effects. Testing reduces the intrinsic motivation of students to learn (Harlen and Crick, 2003). Most studies distinguish between learning goals and performance goals, and note that the introduction of testing realigns the objective towards performing well on the test, as opposed to learning the material. This emphasis on performance goals discourages "deep learning" and promotes rote memorization (Kellaghan et al., 1996). Madaus and Clarke (2001) assert that standardized tests increase the high school dropout rates among minority populations in part because of motivational issues. Related to motivation, Ames (1992) notes that the classroom structure leads to social comparisons between peers, further harming students' self-perception as learners. Students may also suffer from testing fatigue, since test-taking is taxing on both body and mind.

Tests take time to administer, crowding out other instructional activities. Testing also has the tendency to encourage teaching to the test, as teachers focus on only the content

covered in the test, thereby reducing the breadth of materials students learn (Prodromou, 1995; Firestone et al., 1998).[3] Other than class time, resources such as space and labor may be diverted towards test administration.

Test anxiety—the fear of a poor assessment result—is another oft-cited reason behind negative effects. Psychologists have long theorized that anxiety and performance have an inverted "U" shape relationship (Yerkes and Dodson, 1908). Initially, anxiety motivates effort and improves performance, but past a certain point, outcomes deteriorate as the additional anxiety impairs performance. McDonald (2001) comprehensively surveys research on test anxiety, focusing on school-aged children in non-experimental settings. While anxiety varies with age and sex, he finds strong evidence that supports a negative correlation between test anxiety and test performance, although he admits that almost all of the results in the literature reviewed lack a causal interpretation. These more-recent results are largely consistent with previous reviews of the literature (see Hembree, 1988), though there are also correlational studies that have found no relationship between testing and performance.[4]

This paper addresses issues surrounding these conflicting results—small-scale experimental studies that usually find positive testing effects versus mostly correlational studies that often find negative effects—and contributes to the literature on testing effects in three ways. Firstly, compared to previous studies that often relied on small samples in specialized settings, my leveraging of the NAEP as a natural experiment allows me to measure the testing effect on a much larger scale. The NAEP sample itself is large, and the inclusion of control schools brings the total size of the panel to tens of thousands of schools. Secondly, NAEP is conducted in a real-world setting rather than an artificial laboratory setting. In Section 6, I argue that NAEP testing procedures are in fact very similar to what students commonly experience in many state-wide standardized tests. Given the large scale and real-world setting, the results have external validity and are generalizable to the overall population of students

[3]More questionable behavior by administrators to "game the system" has also been documented by Heilig and Darling-Hammond (2008).

[4]See Hart et al. (2015), which also comprehensively documents the extent of testing in a number of school districts in the United States.

in the United States. Finally, though not experimental, a critical advantage of this novel identification strategy is that it is quasi-experimental. In terms of internal validity, because of the random sampling in NAEP, my estimates of the testing effect are unbiased, causal, and as credible as those derived from experimental settings.

3 Background on NAEP

The National Assessment of Educational Progress, or NAEP, is a nationally-representative survey of students in the United States that assesses students in various subject areas. The set of subjects in which students are tested rotate each year; in 2011 and 2013 in particular, math and reading tests were administered. Using the taxonomy described in the previous section, the NAEP tests have the following characteristics.

Questions The NAEP consists of both multiple-choice and constructed response questions. Each student is presented with only a random subset of all possible questions in customized booklets. Furthermore, each student is tested in only one subject, even though each NAEP testing round contains several subjects. This is done to maximize content coverage while minimizing the time burden on students.

Standardization The universe of all possible NAEP test questions with which students are tested is standardized across the entire United States. Testing is administered in a uniform manner across all schools in all states.

Target Group The NAEP randomly selects and tests only a representative sample of students rather than the entire population. Moreover, only students in grades 4, 8, and 12 are tested. There is no self-selection into the sample by participants, although selected students can opt out.

Purpose The NAEP is a purely summative assessment without stakes, used by the federal Department of Education as a common metric for educational progress. Testing is

conducted by outside contractors, with little participation from teachers. Results for individual students are not released; thus, there is no formative feedback derived from the test.

The NAEP uses a stratified random sampling procedure to select schools within each state.[5] This procedure is repeated for each of the grades tested. The Common Core of Data (CCD), a comprehensive directory of all public schools in the country, serves as the sampling frame.[6] The CCD dataset is partitioned by state, and sampling is conducted in each state separately. Within each state, the list of schools is divided into strata based on school characteristics such as location, racial and ethnic composition, and state-based achievement scores. Schools are then randomly selected within each stratum, where the probability of selection depends on the size of that school's grade relative to the state's student population in that grade. This stratification ensures that the sample of schools is nationally representative and has good coverage of schools with different characteristics.

After the schools have been identified and confirmed to be eligible for NAEP participation, a list of all students in the grade to be tested is drawn up at each school. About 60 students are then randomly chosen with equal probability from this school list for testing. Each selected student is assigned a single subject area to be tested in. The accuracy of individual student demographic information is verified with the school.

Testing takes place from late January to early March, and is conducted in a standardized fashion by state coordinators employed by NAEP. Selected students are brought to a testing area provided by the school, and are seated in a prescribed order. Coordinators then distribute booklets of questions, also bundled in a predetermined order such that no two students sitting next to one another receive the same subject. Each booklet contains approximately 30 to 40 questions on the subject being tested, as well as background infor-

[5]The sampling procedure described here is for the NAEP State Assessments. This is the process for the math and reading tests conducted every other year, which is the treatment in the analysis to follow.

[6]Public schools, as referred to in this paper, include publicly-funded charter schools. A similar procedure is done for private schools. However, the analysis to follow will focus only on public schools because of data availability for the subsequent state-wide standardized tests.

mation questions. Testing commences and the students are given 60 minutes to complete the booklet. Students with disabilities and English language learners are given accommodations where possible. At any point, the student may stop taking the test, in effect opting out. Students may also leave questions unanswered. No penalty is assessed for not fully completing the assessment. An entire NAEP testing session typically takes 90 minutes.

Sample attrition is possible in several ways. When informed of their selection, a school may decline to participate. A replacement school with similar characteristics to the school that dropped out is then drawn from the CCD state list as a substitute. Refusal to participate by a public school is very uncommon—for each state in each NAEP round, only one or two public schools, if any, ever declined.

When students are selected within the school, their parents receive a letter notifying them of their child's selection. Since participation is voluntary, a parent can choose to opt their student out of NAEP testing. A student may also happen to be absent on the day of testing due to illness or other circumstances.[7] Student-weighted participation rates across all states and years are almost always above 90 percent.

Because the subsequent analyses are conducted at the school level, the type of non-response bias that is of concern is that arising from schools declining to participate. I do not consider this to be a big issue, given the extremely high response rates for schools.

There is one instance of oversampling in the NAEP. Certain large urban school districts are selected as part of the Trial Urban District Assessment (TUDA).[8] Schools in TUDA districts are oversampled so that TUDA-district-level results can be calculated more precisely and compared to national and state results. The TUDA represents an augmentation to the

[7]Make-up sessions on alternate days are sometimes conducted.

[8]As of 2013, the TUDA districts are: Albuquerque Public Schools, Atlanta Public Schools, Austin Independent School District, Baltimore City Public Schools, Boston Public Schools, Charlotte-Mecklenburg Schools, Chicago Public Schools, Cleveland Metropolitan School District, Dallas Independent School District, Detroit Public Schools, District of Columbia Public Schools, Fresno Unified School District, Hillsborough County (FL) Public Schools, Houston Independent School District, Jefferson County Public Schools (Louisville, KY), Los Angeles Unified School District, Miami-Dade County Public Schools, Milwaukee Public Schools, New York City Department of Education, San Diego Unified School District, and the School District of Philadelphia. The number of TUDA districts has grown from 6 in 2002 (when TUDA first began) to 21 in 2013.

main NAEP, and follows administration procedures identical to the main study. Its sampling process mirrors that of the main NAEP, save for the probability of schools being chosen to participate. Schools in TUDA districts have dissimilar characteristics compared to the average school in the overall NAEP sample. However, this will not matter for the analysis because selection into treatment group is still random and exogenous given the sampling design. Moreover, regression specifications will include a host of school-level characteristics as controls as well as district fixed effects, which will capture any impact resulting from TUDA.

After NAEP testing finishes for that year, state-wide standardized tests are administered in most states. These are usually held between April and June and the exact timing varies by state. I exclude states that conduct such testing in the fall (pre-treatment) or that have wide testing windows which may coincide with NAEP. These standardized tests are mandated by NCLB and state-level legislation, with each state setting its own standards and procedures. Participation in NAEP does not preclude or excuse a student from taking these state-wide tests. Unlike NAEP, in which only a random subset of students in a random subset of schools take part, state-wide tests are administered to all students in all public schools. The school-level scores from these tests serve as outcomes for comparison between NAEP treated schools and non-NAEP control schools.

4 Empirical Findings

4.1 Data

For the analysis, the data come from multiple sources. I use three data sources from the National Center for Education Statistics (NCES). All references to years refer to the school year ending year; that is, 2011 refers to the school year 2010-2011.

The first data source is the restricted-access NAEP data, which details the performance

results of students participating in NAEP testing.[9] In the original data, each observation is a student-year-grade. For this analysis, these observations are aggregated into school-year-grade observations; thus, variable values represent the average of a grade at a school in a particular year. Showing up in this dataset implies the school received the NAEP treatment for that grade in that year. If the school is missing from the data, then it is in the control group. Other variables of note include the number of students tested in each subject and their NAEP scaled scores.[10] I only use NAEP data from 2011 and 2013 for grades 4 and 8. For these years and grades, subject tests in reading and math are always administered, while science tests are administered only to grade 8 students in 2011.[11]

The second data source is the Common Core of Data (CCD) directory of all *public* schools in the United States.[12] The dataset is reshaped so that each observation is a school-year-grade. The data contain variables detailing enrollment at each school and grade level, broken down by race/ethnicity and sex, as well as the number of students participating in free and reduced lunch programs. School-level characteristics are also available, including a school district identifier, locale type (city, suburb, town, or rural), whether the school is a charter school, and the span of grades served by the school. These CCD data are used to derive the control variables in the regressions.

The third data source is the restricted-access EDFacts data on state-wide standardized testing from the various state education departments collated by the NCES.[13] Each observation is a school-year-grade containing achievement results from each state's state-wide standardized test. From this dataset, I use only data from grades 4 and 8 in years 2011 and 2013, and lagged data from grades 3 and 7 in years 2010 and 2012. Achievement results are expressed as percentages of students at each school that fall into ordinal performance lev-

[9] http://nces.ed.gov/nationsreportcard/

[10] Scaled scores range around 100 to 500. For each student, five plausible values are reported. See Mislevy et al. (1992) for a description of the plausible values methodology.

[11] While NAEP tests in other subjects such as history and the arts are administered in even-numbered years, I ignore these other subjects (and hence ignore even-numbered years) because reading and math are the primary outcome measures from the state-wide standardized tests data.

[12] http://nces.ed.gov/ccd/

[13] http://www2.ed.gov/about/inits/ed/edfacts/index.html

els (e.g. "advanced", "basic", "acceptable"). Each state separately defines what constitutes academic standards deserving classification into any performance level. In the 2012 and 2013 data, each state has 3 to 5 performance levels. In 2010 and 2011, all states are observed to have only two overarching levels ("proficient" and "not proficient"), which are generated by NCES by mapping the various state performance levels into these two overarching levels.

I use these percentages from the EDFacts data to create school-level average standardized score variables in math and reading. This procedure is described in detail in Supplemental Appendix A and assumes student performance is normally distributed. The procedure uses the NAEP data mentioned previously to ensure that the generated scores are comparable across states. This standardized score in each subject measures the average performance of a grade at a school in a particular year. Scores are calculated only for public schools, as private schools are not required to participate in state-wide standardized tests. Though each observation in the data is a school-year-grade, the standardized score itself is normalized into a Z-score such that it has a mean of 0 and a standard deviation of 1 over the *national* distribution of achievement at the *student* level. As such, effect estimates should be interpreted as the impact of treatment on state-wide standardized test scores as measured in standard deviation units (σ) along the national *student* achievement distribution.

For each subject, standardized scores from grades 4 and 8 in years 2011 and 2013 serve as my outcome of interest. The corresponding lagged standardized scores from the previous grades 3 and 7 in years 2010 and 2012, from the same cohort within the same school, are appended to each observation as additional variables to be used as regression covariates.

The three data sources described above are merged at the school-year-grade level. From the combined dataset, I exclude 7 states that conduct state-wide standardized testing in the fall, prior to NAEP treatment.[14] I also exclude Hawaii because it has a wide testing window spanning October through May, where the precise timing of administering the state test is up to the school. I further drop Alabama and Wyoming because of missing data issues. This

[14]These states are Maine, Michigan, New Hampshire, North Dakota, Rhode Island, Vermont, and Wisconsin.

leaves 40 states and the District of Columbia in the analysis sample.

Table 2 shows summary statistics calculated over schools for variables from the combined dataset. The first two columns display statistics at the grade 4 level in years 2011 and 2013 respectively. The third and fourth columns display statistics at the grade 8 level in years 2011 and 2013 respectively. On average, about 13% of schools containing grade 4 are NAEP-tested, while around 21% to 23% of schools containing grade 8 are NAEP-tested. The treatment intensity variable measures the proportion of students in the grade at a treated school who are selected to sit for the NAEP that year. This varies across schools because around the same number of students (60) is selected to take the NAEP at each school, regardless of grade size.[15] Thus, at smaller schools, treatment intensity will be greater. The summary statistics for treatment intensity in Table 2 are calculated over treated schools only; however, in the analysis to follow, treatment intensity is coded as 0 at control schools.

Note that for standardized scores in math and reading, the means and standard deviations are not precisely 0 and 1, despite their being Z-scores. There are two reasons for this. First, the scores are constructed such that the normalization is at the student-level over the national distribution of achievement. The statistics in Table 2, however, are calculated over school-level averages of these standardized scores, so their standard deviations will be smaller than 1. Second, the standardized scores are constructed using data from all 50 states and the District of Columbia, but the analysis sample excludes 10 of these states. So while the means are all close to 0, they are slightly off.

In total, the analysis sample contains over 40,000 schools containing grade 4, and over 20,000 schools containing grade 8, for each year being analyzed. These and all subsequent reported sample sizes (including number of clusters) are rounded to the nearest 10 schools

[15]NAEP-tested schools with fewer than 60 students in the tested grade are the exception. In this case, all students in the tested grade sit for the NAEP. When calculating treatment intensity, a few observations were greater than 1, and these were topcoded to 1. This is because the numerator (number of students tested by NAEP) comes from the NAEP dataset measured in the spring during NAEP testing, while the denominator (enorollment in that grade at that school) comes from the CCD dataset measured at the beginning of fall before NAEP testing. Some students may have joined the grade in the intervening time. This is a minor issue as very few schools were topcoded.

Table 2: Summary Statistics

	Grade 4		Grade 8	
Variable	2011	2013	2011	2013
NAEP-Tested	0.131	0.128	0.229	0.209
	(0.338)	(0.334)	(0.421)	(0.407)
Treatment Intensity	0.814	0.769	0.546	0.444
(over tested schools only)	(0.200)	(0.220)	(0.293)	(0.297)
Math Score	0.004	-0.024	-0.038	-0.064
	(0.222)	(0.311)	(0.314)	(0.315)
Lagged Math Score	0.005	0.033	-0.021	0.038
	(0.244)	(0.335)	(0.322)	(0.347)
Reading Score	0.015	-0.004	-0.018	-0.040
	(0.207)	(0.288)	(0.229)	(0.286)
Lagged Reading Score	-0.032	0.012	-0.033	-0.034
	(0.252)	(0.361)	(0.258)	(0.319)
Grade % White	0.522	0.507	0.554	0.541
	(0.344)	(0.344)	(0.35)	(0.351)
Grade % Black	0.162	0.157	0.169	0.167
	(0.252)	(0.248)	(0.265)	(0.264)
Grade % Hispanic	0.230	0.243	0.197	0.207
	(0.279)	(0.283)	(0.263)	(0.265)
Grade % Native American	0.018	0.018	0.027	0.027
	(0.089)	(0.092)	(0.117)	(0.119)
Grade % Asian	0.041	0.043	0.030	0.031
	(0.085)	(0.090)	(0.072)	(0.074)
Grade % Hawaii / Pacific	0.002	0.002	0.002	0.002
	(0.012)	(0.010)	(0.013)	(0.010)
Grade % Multiracial	0.026	0.030	0.020	0.024
	(0.046)	(0.043)	(0.048)	(0.047)
Grade % Boys	0.515	0.514	0.521	0.517
	(0.091)	(0.090)	(0.121)	(0.114)
Grade Size	73.64	74.07	140.19	143.13
	(41.96)	(41.98)	(134.45)	(134.56)
School Size	474.48	485.55	511.85	529.33
	(252.17)	(280.30)	(381.00)	(416.45)
School % Free / Red. Lunch	0.533	0.561	0.521	0.547
	(0.285)	(0.283)	(0.276)	(0.272)
Charter School	0.048	0.058	0.088	0.102
	(0.215)	(0.233)	(0.283)	(0.303)
Locale:				
City	0.303	0.309	0.265	0.274
Suburb	0.304	0.348	0.235	0.270
Town	0.100	0.096	0.120	0.118
Rural	0.293	0.247	0.380	0.338
N (Schools)	43550	41020	23090	21770

Notes: Means and standard deviations (in parentheses) calculated over schools, the unit of observation. Sample sizes rounded to the nearest 10.

to comply with NCES data confidentiality rules.

4.2 Regression Analysis

With these data, I estimate the treatment effect of NAEP testing separately for each subject (math and reading), grade (4 and 8) and year (2011 and 2013), utilizing the value-added regression model

$$y_{sd} = \beta D_{sd} + \delta y_{sd}^{lagged} + X_{sd}\gamma + \alpha_d + \varepsilon_{sd} \tag{1}$$

where

- y_{sd} is the standardized score derived from state-wide standardized test results in math or reading at school s in district d,
- D_{sd} is the NAEP treatment indicator, equaling 1 if school s in district d is randomly selected for NAEP testing that year,
- y_{sd}^{lagged} is the lagged standardized score in math or reading at school s in district d (that is, that cohort's score from the previous year, when they were one grade lower),
- X_{sd} is a set of controls from the CCD,
- α_d are district fixed effects, and
- ε_{sd} is an error term.

The full set of controls in X_{sd} is: proportions of Black / Hispanic / Native American / Asian / Hawaiian and Pacific Islander / Multiracial students (White is omitted) in the specific grade, proportion of boys in the specific grade, number of students in the specific grade, proportion of students in free and reduced lunch programs at the school, indicator for charter school, and category dummies for school locale being a suburb / town / rural (city is omitted).[16] For all the analyses presented in this section, checks for robustness to using alternative sets

[16] I use grade-level covariates where available; where unavailable, I use school-level covariates.

of controls are carried out in Supplemental Appendix B; the findings there reinforce the validity of the main results below.

The coefficient estimate of β in regression (1) measures the treatment effect (in σ units of standardized score) of NAEP testing on state-wide standardized test score gains. Under the value-added approach, the treatment effect is the difference in gains in test score from one grade to the next between the treatment and control groups. Since both groups take the subsequent state-wide standardized tests which measures the outcome y_{sd}, the treatment effect is the marginal effect of the additional NAEP test experienced by the treatment group only. Identification hinges on the fact that treatment status D_{sd} is randomly assigned in the NAEP sampling process, and therefore uncorrelated with the error term ε_{sd}, conditional on covariates X_{sd}. The inclusion of these covariates in the regression is essential because the stratified sampling procedure involves some of these variables. Overall, this identification strategy ensures that estimates of β do not suffer from omitted variable bias and can be interpreted as causal.

Table 3 reports the effect of NAEP testing on state-wide standardized test scores. Panel (A) reports results with math scores as the dependent variable y_{sd} in the regression (1), while Panel (B) reports results with reading scores as the dependent variable. Columns (1) and (2) correspond to regressions using grade 4 data, while columns (3) and (4) correspond to regressions using grade 8 data. Regressions are run separately for year 2011 (columns (1) and (3)) and year 2013 (columns (2) and (4)).

The results suggest that NAEP testing has virtually no effect on subsequent state-wide standardized test scores. All point estimates are close to zero, with no magnitude exceeding 0.006σ. As a result of these small magnitudes, many of the estimates are not statistically significant. Even the ones that are statistically significant indicate minute negative impacts. For example, the largest-magnitude statistically-significant treatment effect is observed for grade 4 reading scores in 2011. This estimate suggests that students at schools selected for NAEP testing score on average 0.0059σ lower in subsequent state-wide standardized tests

Table 3: Effect of NAEP Testing on Standardized Scores

Panel (A): Math	(1)	(2)	(3)	(4)
Dep. Var.:	Grade 4		Grade 8	
Math Score	2011	2013	2011	2013
NAEP-Tested	-0.0059***	-0.0001	-0.0020	0.0040
	(0.0021)	(0.0028)	(0.0039)	(0.0036)
Lagged Math Score	0.545***	0.612***	0.703***	0.730***
	(0.014)	(0.011)	(0.015)	(0.016)
Controls	Yes	Yes	Yes	Yes
District FEs	Yes	Yes	Yes	Yes
N (Schools)	43490	41020	22950	21710
Clusters (Districts)	11390	11000	11540	11130
R-Square	0.433	0.536	0.604	0.661

Panel (B): Reading	(1)	(2)	(3)	(4)
Dep. Var.:	Grade 4		Grade 8	
Reading Score	2011	2013	2011	2013
NAEP-Tested	-0.0057***	-0.0058**	-0.0019	-0.0010
	(0.0016)	(0.0024)	(0.0029)	(0.0033)
Lagged Reading Score	0.557***	0.571***	0.653***	0.743***
	(0.016)	(0.015)	(0.017)	(0.019)
Controls	Yes	Yes	Yes	Yes
District FEs	Yes	Yes	Yes	Yes
N (Schools)	43500	41010	23050	21750
Clusters (Districts)	11390	11000	11560	11140
R-Square	0.525	0.661	0.643	0.707

Significance Levels: *** = 1% ; ** = 5%; * = 10%.

Notes: Each column within a panel reports results for a separate regression with the dependent variable being school-average standardized scores for a specific grade and year. Standard errors clustered at the district level are in parentheses. All regressions include the "NAEP-Tested" treatment indicator, lagged standardized scores, controls (estimates not shown), and district fixed effects. The full set of controls is: proportions of Black / Hispanic / Native American / Asian / Hawaiian and Pacific Islander / Multiracial students (White is omitted) in the specific grade, proportion of boys in the specific grade, number of students in the specific grade, proportion of students in free and reduced lunch programs at the school, indicator for carter school, and category dummies for school locale being a suburb / town / rural (city is omitted). Sample sizes rounded to the nearest 10.

compared to students at control schools. Overall, these estimates suggest that the effects of testing are negligible.

How economically meaningful—or rather economically trivial—are these treatment effect sizes? Focusing only on absolute values, the largest magnitude estimated was approximately 0.006σ. For comparison, consider the results from Angrist and Lavy (1999). Using cutoffs in Maimonides' Rule in their identification strategy, they estimate that increasing class size by one student has a negative effect of 0.022σ.[17] This implies that the 0.006σ effect of the hour-long NAEP test is similar to increasing class size by about one-quarter of a student. Such magnitudes indeed seem relatively inconsequential. Another way to think about this is that one would need to administer four standardized tests similar to the NAEP treatment before having the equivalent effect of increasing class size by one student.

The β coefficients estimated above measure treatment effects regardless of treatment intensity. However, as noted previously, only a subset of students at schools selected for NAEP is administered the test, with treatment intensity varying from school to school. In the specifications above, the NAEP-tested indicator D_{stg} equals 1 regardless of the proportion of students tested. In order to estimate treatment effects on both the intensive and extensive margins, I add the NAEP treatment intensity measure (a proportion between 0 and 1) as an extra covariate in regression (1). This measure is always coded 0 for control schools. Treatment intensity is exogenous conditional on grade size, which is included as a control variable in X_{stg}. Thus, in this augmented specification, the coefficient on the NAEP-tested indicator measures the testing effect on the extensive margin (the effect from NAEP coordinators showing up at the school to theoretically administer the test to zero students) while the coefficient on the treatment intensity measures the testing effect on the intensive margin (the effect of "scaling up" the proportion of students being tested).[18]

[17] See their p. 567 and footnote 22. This number is calculated by first dividing -0.275 by 7.7, which gives the effect of increasing class size by one student in the distribution of class average test scores. This effect size is then scaled down using the ratio $\frac{0.18}{0.29}$ to arrive at the effect of increasing class size by one student in the overall student distribution.

[18] The linearity behind this specification assumes a constant "returns to scale" in the effect of testing.

Estimates from the augmented specification are presented in Table (4), arranged in the same 2-panel-4-column format as before. These estimates present a more nuanced picture. On the extensive margin, the estimates are all negative and seven out of eight are statistically significant at least at the 10% level, with four of the eight estimates being significant at the 5% level. The estimates range from -0.009σ to -0.031σ—magnitudes much larger than before, but still relatively small compared to the literature. This suggests that having NAEP coordinators just show up at the school (without actually testing any students) has a negative impact on subsequent state-wide test results.

On the intensive margin, the estimates are all positive, with five of the eight estimates being statistically significant at the 10% level. These estimates range from 0.013σ to 0.045σ, suggesting that scaling up the proportion of students being tested actually has a positive impact on subsequent state-wide test results. Note that for each intensive-margin estimate, the magnitude somewhat mirrors the corresponding extensive-margin estimate found for that particular subject-grade-year. For instance, looking at math scores for grade 4 students in 2011, the negative effect on the extensive margin is -0.031 while the positive effect on the intensive margin is 0.031.

The policy relevant treatment effect is actually the intensive- and extensive-margin estimates added together. This corresponds to the treatment effect at treatment intensity equaling 1. This is the effect of interest to policy-makers because were one additional NAEP-like standardized test administered, it would presumably be administered to all students. I calculate this effect at intensity 1 for each regression specification using the delta method and present these results in a row towards the bottom of each panel. Only two of the eight effect estimates are statistically significant, while the remaining six estimates are insignificant and close to zero. The two statistically significant estimates for grade 8 math and reading scores in 2013 are positive and statistically significant at the 5% level, with magnitudes 0.030 and 0.026 respectively. This would suggest that for grade 8 students in 2013, fully implementing the NAEP test at a school actually has a positive impact on subsequent state-wide test scores,

Table 4: Intensive and Extensive Margin Effects

Panel (A): Math	(1)	(2)	(3)	(4)
Dep. Var.:	Grade 4		Grade 8	
Math Score	2011	2013	2011	2013
NAEP-Tested	-0.031***	-0.013	-0.014*	-0.014**
	(0.007)	(0.009)	(0.008)	(0.006)
Treatment Intensity	0.031***	0.018	0.024*	0.045***
	(0.009)	(0.012)	(0.013)	(0.013)
Lagged Math Score	0.544***	0.612***	0.703***	0.729***
	(0.014)	(0.011)	(0.015)	(0.016)
Controls	Yes	Yes	Yes	Yes
District FEs	Yes	Yes	Yes	Yes
Effect at Intensity=1	0.0001	0.004	0.010	0.030***
	(0.003)	(0.004)	(0.008)	(0.009)
N (Schools)	43490	41020	22950	21710
Clusters (Districts)	11390	11000	11540	11130
R-Square	0.433	0.536	0.604	0.661

Panel (B): Reading	(1)	(2)	(3)	(4)
Dep. Var.:	Grade 4		Grade 8	
Reading Scores	2011	2013	2011	2013
NAEP-Tested	-0.016***	-0.015*	-0.009*	-0.020***
	(0.006)	(0.009)	(0.005)	(0.005)
Treatment Intensity	0.013*	0.013	0.014	0.045***
	(0.007)	(0.011)	(0.011)	(0.015)
Lagged Reading Score	0.557***	0.571***	0.652***	0.741***
	(0.016)	(0.015)	(0.017)	(0.019)
Controls	Yes	Yes	Yes	Yes
District FEs	Yes	Yes	Yes	Yes
Effect at Intensity=1	-0.003	-0.003	0.005	0.026**
	(0.002)	(0.004)	(0.007)	(0.011)
N (Schools)	43500	41010	23050	21750
Clusters (Districts)	11390	11000	11560	11140
R-Square	0.525	0.661	0.643	0.708

Significance Levels: *** = 1% ; ** = 5%; * = 10%.

Notes: Each column within a panel reports results for a separate regression with the dependent variable being school-average standardized scores for a specific grade and year. Standard errors clustered at the district level are in parentheses. All regressions include the "NAEP-Tested" treatment indicator, treatment intensity as measured by proportion of students tested for NAEP, lagged standardized scores, controls (estimates not shown), and district fixed effects. The full set of controls is: proportions of Black / Hispanic / Native American / Asian / Hawaiian and Pacific Islander / Multiracial students (White is omitted) in the specific grade, proportion of boys in the specific grade, number of students in the specific grade, proportion of students in free and reduced lunch programs at the school, indicator for carter school, and category dummies for school locale being a suburb / town / rural (city is omitted). Sample sizes rounded to the nearest 10.

approximately equivalent to reducing class size by one student. Overall though, the majority of effect estimates at intensity 1 suggest that administering a NAEP-like standardized test to all students at a school has an inconsequential effect.

Nevertheless, separating out the effects on the intensive and extensive margins provides a way to think about the mechanisms behind the treatment effects. I later discuss different possible channels in Section 5, but before doing so, I first consider heterogeneity in treatment effect sizes by sex and check for random assignment of treatment status in the following two subsections.

4.3 Heterogeneity by Sex

The treatment effect of NAEP testing may be heterogeneous between the sexes. For instance, girls and boys could respond differently to competitive environments (Niederle and Vesterlund, 2010). In addition to the overall average test results at the school level for each year and grade, the data also reports average results by sex at the school level (i.e. the average among all girls or all boys within a school). Using these finer measures, I repeat the regressions in Tables 3 and 4 to estimate treatment effects for girls and boys separately.

The corresponding results from specifications with only the NAEP-tested indicator are presented in Table 5 for girls and Table 6 for boys. For both sexes, all the point estimates of the coefficient on the NAEP-tested indicator are close to zero. Consequently, many of the estimates are not statistically significant. For girls, the largest-magnitude treatment effect observed is only 0.0054σ. For boys, the largest-magnitude observed is only 0.0088σ. These results are similar to those found in the overall results, suggesting that NAEP testing has a negligible effect on the subsequent state-wide test results of both boys and girls.

The corresponding results from specifications including both the NAEP-tested indicator and the treatment intensity measure are presented in Table 7 for girls and Table 8 for boys. A similar pattern emerges for both sexes, with the coefficient estimates on the NAEP-tested indicator being negative and those on the treatment intensity measure being positive. Again,

Table 5: Effect of NAEP Testing on Standardized Scores (Girls Only)

Panel (A): Math	(1)	(2)	(3)	(4)
Dep. Var.:	Grade 4		Grade 8	
Math Score	2011	2013	2011	2013
NAEP-Tested	-0.0054**	-0.0011	-0.004	0.0051
	(0.0023)	(0.0031)	(0.0041)	(0.0041)
Lagged Math Score	0.506***	0.586***	0.637***	0.702***
	(0.013)	(0.011)	(0.018)	(0.019)
Controls	Yes	Yes	Yes	Yes
District FEs	Yes	Yes	Yes	Yes
N (Schools)	42940	40620	22160	21160
Clusters (Districts)	11230	10860	11390	11010
R-Square	0.381	0.490	0.519	0.605

Panel (B): Reading	(1)	(2)	(3)	(4)
Dep. Var.:	Grade 4		Grade 8	
Reading Score	2011	2013	2011	2013
NAEP-Tested	-0.0054***	-0.0026	-0.0039	0.0034
	(0.0018)	(0.0028)	(0.0032)	(0.0037)
Lagged Reading Score	0.524***	0.536***	0.578***	0.691***
	(0.014)	(0.015)	(0.021)	(0.021)
Controls	Yes	Yes	Yes	Yes
District FEs	Yes	Yes	Yes	Yes
N (Schools)	43000	40620	22280	21210
Clusters (Districts)	11240	10860	11420	11040
R-Square	0.469	0.606	0.529	0.651

Significance Levels: *** = 1% ; ** = 5%; * = 10%.

Notes: Each column within a panel reports results for a separate regression with the dependent variable being school-average standardized scores (averaged over girls only) for a specific grade and year. Standard errors clustered at the district level are in parentheses. All regressions include the "NAEP-Tested" treatment indicator, lagged standardized scores, controls (estimates not shown), and district fixed effects. The full set of controls is: proportions of Black / Hispanic / Native American / Asian / Hawaiian and Pacific Islander / Multiracial students (White is omitted) in the specific grade, proportion of boys in the specific grade, number of students in the specific grade, proportion of students in free and reduced lunch programs at the school, indicator for carter school, and category dummies for school locale being a suburb / town / rural (city is omitted). Sample sizes rounded to the nearest 10.

Table 6: Effect of NAEP Testing on Standardized Scores (Boys Only)

Panel (A): Math	(1)	(2)	(3)	(4)
Dep. Var.:	Grade 4		Grade 8	
Math Score	2011	2013	2011	2013
NAEP-Tested	-0.0070***	0.0006	-0.0029	0.0015
	(0.0023)	(0.0031)	(0.0043)	(0.0036)
Lagged Math Score	0.504***	0.599***	0.656***	0.714***
	(0.015)	(0.011)	(0.016)	(0.014)
Controls	Yes	Yes	Yes	Yes
District FEs	Yes	Yes	Yes	Yes
N (Schools)	43190	40790	22560	21420
Clusters (Districts)	11280	10910	11450	11040
R-Square	0.383	0.511	0.544	0.644

Panel (B): Reading	(1)	(2)	(3)	(4)
Dep. Var.:	Grade 4		Grade 8	
Reading Score	2011	2013	2011	2013
NAEP-Tested	-0.0058***	-0.0088***	-0.0023	-0.0030
	(0.0018)	(0.0028)	(0.0033)	(0.0036)
Lagged Reading Score	0.530***	0.569***	0.612***	0.706***
	(0.014)	(0.015)	(0.020)	(0.021)
Controls	Yes	Yes	Yes	Yes
District FEs	Yes	Yes	Yes	Yes
N (Schools)	43170	40790	22650	21460
Clusters (Districts)	11270	10910	11450	11060
R-Square	0.477	0.603	0.574	0.650

Significance Levels: *** = 1% ; ** = 5%; * = 10%.

Notes: Each column within a panel reports results for a separate regression with the dependent variable being school-average standardized scores (averaged over boys only) for a specific grade and year. Standard errors clustered at the district level are in parentheses. All regressions include the "NAEP-Tested" treatment indicator, lagged standardized scores, controls (estimates not shown), and district fixed effects. The full set of controls is: proportions of Black / Hispanic / Native American / Asian / Hawaiian and Pacific Islander / Multiracial students (White is omitted) in the specific grade, proportion of boys in the specific grade, number of students in the specific grade, proportion of students in free and reduced lunch programs at the school, indicator for carter school, and category dummies for school locale being a suburb / town / rural (city is omitted). Sample sizes rounded to the nearest 10.

the policy-relevant treatment effects at intensity 1 are the ones of most interest. As before, almost all of these are small in magnitude and not statistically significant at the 5% level. Once more, the main exception is the treatment effects for grade 8 students in 2013. For grade 8 girls in 2013, the treatment effects at intensity 1 are 0.036σ for math and 0.028σ for reading, with both being statistically significant at the 5% level. For grade 8 boys in 2013, the treatment effects at intensity 1 are 0.021σ for both math and reading, with the latter being only statistically significant at the 10% level. This suggests that any positive effect from testing seen for this particular 2013 grade 8 sample is driven mainly by girls. Aside from this irregular grade-year though, there is little evidence that fully implementing testing at schools as an effect on subsequent state-wide standardized test scores for either girls or boys.

4.4 Checking Random Assignment

I verify the random assignment of treatment status by regressing the NAEP-tested indicator on lagged scores and controls (presumably predetermined variables), using the linear probability regression

$$D_{sd} = \pi y_{sd}^{lagged} + X_{sd}\mu + \alpha_d + \varepsilon_{sd} \tag{2}$$

If NAEP treatment is randomly assigned, then the coefficients π and (the vector) μ should be zero. A caveat to this check for randomness, however, is that variables on the right hand side of regression (2) may be used for stratification in the sampling procedure, so it would not be surprising if a few have an "effect" on treatment D_{sd}.

Table 9 investigates whether treatment is assigned randomly for the math samples. The coefficient estimates of regression (2) are reported for the corresponding four samples in Panel (A) columns (1) through (4) of earlier tables. Table 10 investigates whether treatment is assigned randomly for the reading samples. The coefficient estimates of regression (2) are reported for the corresponding four samples in Panel (B) columns (1) through (4) of earlier

Table 7: Intensive and Extensive Margin Effects (Girls Only)

Panel (A): Math	(1)	(2)	(3)	(4)
Dep. Var.:	Grade 4		Grade 8	
Math Score	2011	2013	2011	2013
NAEP-Tested	-0.020**	-0.015	-0.025***	-0.016**
	(0.008)	(0.009)	(0.008)	(0.007)
Treatment Intensity	0.017*	0.019	0.043***	0.052***
	(0.010)	(0.013)	(0.015)	(0.014)
Lagged Math Score	0.505***	0.585***	0.635***	0.701***
	(0.013)	(0.011)	(0.018)	(0.018)
Controls	Yes	Yes	Yes	Yes
District FEs	Yes	Yes	Yes	Yes
Effect at Intensity=1	-0.002	0.003	0.018**	0.036***
	(0.003)	(0.005)	(0.009)	(0.010)
N (Schools)	42940	40620	22160	21160
Clusters (Districts)	11230	10860	11390	11010
R-Square	0.381	0.490	0.519	0.606

Panel (B): Reading	(1)	(2)	(3)	(4)
Dep. Var.:	Grade 4		Grade 8	
Reading Score	2011	2013	2011	2013
NAEP-Tested	-0.013**	-0.013	-0.009	-0.013**
	(0.006)	(0.010)	(0.006)	(0.006)
Treatment Intensity	0.010	0.013	0.010	0.041***
	(0.008)	(0.013)	(0.013)	(0.016)
Lagged Reading Score	0.524***	0.536***	0.577***	0.690***
	(0.014)	(0.015)	(0.021)	(0.021)
Controls	Yes	Yes	Yes	Yes
District FEs	Yes	Yes	Yes	Yes
Effect at Intensity=1	-0.004	0.001	0.001	0.028**
	(0.002)	(0.004)	(0.008)	(0.012)
N (Schools)	43000	40620	22280	21210
Clusters (Districts)	11240	10860	11420	11040
R-Square	0.469	0.606	0.529	0.651

Significance Levels: *** = 1% ; ** = 5%; * = 10%.

Notes: Each column within a panel reports results for a separate regression with the dependent variable being school-average standardized scores (averaged over girls only) for a specific grade and year. Standard errors clustered at the district level are in parentheses. All regressions include the "NAEP-Tested" treatment indicator, treatment intensity as measured by proportion of students tested for NAEP, lagged standardized scores, controls (estimates not shown), and district fixed effects. The full set of controls is: proportions of Black / Hispanic / Native American / Asian / Hawaiian and Pacific Islander / Multiracial students (White is omitted) in the specific grade, proportion of boys in the specific grade, number of students in the specific grade, proportion of students in free and reduced lunch programs at the school, indicator for carter school, and category dummies for school locale being a suburb / town / rural (city is omitted). Sample sizes rounded to the nearest 10.

Table 8: Intensive and Extensive Margin Effects (Boys Only)

Panel (A): Math	(1)	(2)	(3)	(4)
Dep. Var.:	Grade 4		Grade 8	
Math Score	2011	2013	2011	2013
NAEP-Tested	-0.028***	-0.013	-0.014*	-0.012*
	(0.007)	(0.010)	(0.008)	(0.006)
Treatment Intensity	0.026***	0.017	0.023	0.033**
	(0.009)	(0.014)	(0.016)	(0.014)
Lagged Math Score	0.504***	0.599***	0.655***	0.714***
	(0.015)	(0.011)	(0.016)	(0.014)
Controls	Yes	Yes	Yes	Yes
District FEs	Yes	Yes	Yes	Yes
Effect at Intensity=1	-0.002	0.005	0.009	0.021**
	(0.003)	(0.005)	(0.010)	(0.009)
N (Schools)	43190	40790	22560	21420
Clusters (Districts)	11280	10910	11450	11040
R-Square	0.383	0.511	0.544	0.644

Panel (B): Reading	(1)	(2)	(3)	(4)
Dep. Var.:	Grade 4		Grade 8	
Reading Score	2011	2013	2011	2013
NAEP-Tested	-0.012*	-0.021**	-0.012**	-0.019***
	(0.007)	(0.009)	(0.006)	(0.006)
Treatment Intensity	0.008	0.016	0.020	0.040**
	(0.008)	(0.012)	(0.013)	(0.016)
Lagged Reading Score	0.530***	0.569***	0.611***	0.706***
	(0.014)	(0.015)	(0.020)	(0.021)
Controls	Yes	Yes	Yes	Yes
District FEs	Yes	Yes	Yes	Yes
Effect at Intensity=1	-0.004*	-0.005	0.008	0.021*
	(0.002)	(0.004)	(0.008)	(0.012)
N (Schools)	43170	40790	22650	21460
Clusters (Districts)	11270	10910	11450	11060
R-Square	0.477	0.603	0.574	0.650

Significance Levels: *** = 1% ; ** = 5%; * = 10%.

Notes: Each column within a panel reports results for a separate regression with the dependent variable being school-average standardized scores (averaged over boys only) for a specific grade and year. Standard errors clustered at the district level are in parentheses. All regressions include the "NAEP-Tested" treatment indicator, treatment intensity as measured by proportion of students tested for NAEP, lagged standardized scores, controls (estimates not shown), and district fixed effects. The full set of controls is: proportions of Black / Hispanic / Native American / Asian / Hawaiian and Pacific Islander / Multiracial students (White is omitted) in the specific grade, proportion of boys in the specific grade, number of students in the specific grade, proportion of students in free and reduced lunch programs at the school, indicator for carter school, and category dummies for school locale being a suburb / town / rural (city is omitted). Sample sizes rounded to the nearest 10.

tables.

Almost all coefficient estimates in these two tables are not statistically different from zero; however, there are a few exceptions. First, in all grades and years, a larger grade size increases the probability of a school being selected for NAEP. This is entirely expected though, because for NAEP sampling, the probability of selection depends on the size of that school's grade. Second, of the eight coefficients on the charter school indicator, three are statistically significant at the 5% level, while another three are at the 10% level, with all point estimates indicating that charter schools are less likely to be selected. Lastly, there are a smattering of coefficient estimates which surpass the 5% level of significance, but these do not exhibit any coherent pattern.

Despite these exceptions, the overall pattern of the coefficient estimates in these checks strongly suggests that NAEP treatment is randomized. There is no indication in the technical documentation that charter schools were considered differently, so it would seem that the random sampling procedure just happened to select fewer charter schools for NAEP testing, particularly for both grades in 2013, and grade 4 in 2011. Three other potential explanations for charter schools being under-sampled are 1) several states do not legally allow them to operate; 2) oversampling in TUDA districts which have fewer charter schools; and 3) charter schools tend to be smaller, so have a lower probability of being sampled (i.e. collinearity with grade size). Overall, there is no reason to believe that NAEP coordinators used a non-random sampling procedure, so any observed oversampling of certain subgroups for only certain years or grades is most likely due to chance. Moreover, the same covariates X_{sd} used in these checks are also included as controls in the main value-added specifications (regression (1)), thereby addressing any remaining concerns relating to chance selection on observables.

Table 9: Check for Random Assignment of Math Sample

	(1)	(2)	(3)	(4)
Dep. Var.:	Grade 4		Grade 8	
NAEP-Tested	2011	2013	2011	2013
Lagged Math Score	0.001	0.001	-0.018	0.032*
	(0.009)	(0.009)	(0.016)	(0.018)
Grade % Black	-0.011	0.005	-0.03	-0.011
	(0.014)	(0.017)	(0.025)	(0.025)
Grade % Hispanic	0.001	-0.003	-0.006	0.010
	(0.017)	(0.016)	(0.027)	(0.028)
Grade % Native American	0.077	-0.072	0.202***	-0.008
	(0.058)	(0.056)	(0.073)	(0.062)
Grade % Asian	-0.037	0.027	-0.022	-0.015
	(0.027)	(0.028)	(0.061)	(0.060)
Grade % Hawaii / Pacific	0.327	0.013	0.342	0.103
	(0.207)	(0.189)	(0.233)	(0.338)
Grade % Multiracial	0.005	-0.027	0.067	-0.039
	(0.045)	(0.054)	(0.09)	(0.068)
Grade % Boys	-0.034*	-0.034	-0.037	-0.044*
	(0.02)	(0.023)	(0.025)	(0.026)
Grade Size	0.0011***	0.0010***	0.0010***	0.0010***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
School % Free / Red. Lunch	0.016	-0.005	0.026	0.035
	(0.014)	(0.015)	(0.024)	(0.028)
Charter	-0.020*	-0.020**	-0.028	-0.035*
	(0.012)	(0.010)	(0.018)	(0.019)
Suburb	0.004	0.016**	0.006	-0.002
	(0.007)	(0.006)	(0.013)	(0.014)
Town	0.009	0.0002	-0.025	0.013
	(0.012)	(0.013)	(0.023)	(0.024)
Rural	-0.004	0.003	-0.029*	-0.005
	(0.008)	(0.009)	(0.015)	(0.018)
District FEs	Yes	Yes	Yes	Yes
N (Schools)	43490	41020	22950	21710
Clusters (Districts)	11390	11000	11540	11130
R-Square	0.012	0.010	0.084	0.088

Significance Levels: *** = 1% ; ** = 5%; * = 10%.

Notes: Each column reports results for a separate regression with the dependent variable being the "NAEP-Tested" treatment indicator for a specific grade and year. Standard errors clustered at the district level are in parentheses. All regressions include lagged standardized scores, controls, and district fixed effects. The full set of controls is: proportions of Black / Hispanic / Native American / Asian / Hawaiian and Pacific Islander / Multiracial students (White is omitted) in the specific grade, proportion of boys in the specific grade, number of students in the specific grade, proportion of students in free and reduced lunch programs at the school, indicator for carter school, and category dummies for school locale being a suburb / town / rural (city is omitted). Sample sizes rounded to the nearest 10.

Table 10: Check for Random Assignment of Reading Sample

	(1)	(2)	(3)	(4)
Dep. Var.:	Grade 4		Grade 8	
NAEP-Tested	2011	2013	2011	2013
Lagged Reading Score	-0.003	0.016	-0.008	0.045***
	(0.010)	(0.010)	(0.019)	(0.017)
Grade % Black	-0.013	0.011	-0.035	-0.006
	(0.014)	(0.017)	(0.024)	(0.025)
Grade % Hispanic	0.0002	0.003	-0.013	0.015
	(0.017)	(0.017)	(0.028)	(0.028)
Grade % Native American	0.086	-0.067	0.161**	-0.006
	(0.058)	(0.056)	(0.077)	(0.063)
Grade % Asian	-0.036	0.025	-0.04	-0.015
	(0.027)	(0.028)	(0.06)	(0.059)
Grade % Hawaii / Pacific	0.370*	0.017	0.174	0.033
	(0.213)	(0.189)	(0.238)	(0.329)
Grade % Multiracial	0.007	-0.027	0.063	-0.054
	(0.045)	(0.054)	(0.087)	(0.070)
Grade % Boys	-0.037*	-0.029	-0.014	-0.041
	(0.019)	(0.023)	(0.024)	(0.026)
Grade Size	0.0011***	0.0009***	0.0010***	0.0010***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
School % Free / Red. Lunch	0.015	0.001	0.032	0.041
	(0.014)	(0.016)	(0.024)	(0.029)
Charter	-0.020*	-0.021**	-0.028	-0.038**
	(0.012)	(0.010)	(0.018)	(0.019)
Suburb	0.004	0.017***	0.007	-0.002
	(0.007)	(0.006)	(0.013)	(0.014)
Town	0.009	0.001	-0.026	0.014
	(0.012)	(0.013)	(0.023)	(0.024)
Rural	-0.004	0.004	-0.028*	-0.004
	(0.008)	(0.009)	(0.015)	(0.018)
District FEs	Yes	Yes	Yes	Yes
N (Schools)	43500	41010	23050	21750
Clusters (Districts)	11390	11000	11560	11140
R-Square	0.012	0.010	0.085	0.090

Significance Levels: *** = 1% ; ** = 5%; * = 10%.

Notes: Each column reports results for a separate regression with the dependent variable being the "NAEP-Tested" treatment indicator for a specific grade and year. Standard errors clustered at the district level are in parentheses. All regressions include lagged standardized scores, controls, and district fixed effects. The full set of controls is: proportions of Black / Hispanic / Native American / Asian / Hawaiian and Pacific Islander / Multiracial students (White is omitted) in the specific grade, proportion of boys in the specific grade, number of students in the specific grade, proportion of students in free and reduced lunch programs at the school, indicator for carter school, and category dummies for school locale being a suburb / town / rural (city is omitted). Sample sizes rounded to the nearest 10.

Table 11: Possible Mechanisms of Testing Effect

	Mechanism	Effect	Margin	Assessment of Mechanism
i)	Formative Learning	Positive	Extensive	Unlikely; NAEP purely summative
		Positive	Intensive	Unlikely; requires student initiative
ii)	Stretched Resources	Negative	Extensive	Possible
		Negative	Intensive	Inconsistent with results
iii)	Perceived Stakes	Positive	Extensive	Inconsistent with results; unlikely
iv)	Intrinsic Motivation	Negative	Intensive	Inconsistent with results
v)	Test Anxiety	Positive	Intensive	Possible
		Negative	Intensive	Inconsistent with results
vi)	Test Fatigue	Negative	Intensive	Inconsistent with results
vii)	Learning by Doing	Positive	Intensive	Possible

5 Discussion

Considering the results presented in Tables 3 and 4, what are some possible mechanisms for explaining the effect of testing? That is, what is the channel through which NAEP treatment acts that would have an effect on subsequent state-wide standardized testing outcomes? To answer this, I explore various possible mechanisms, discerning for each whether theory posits a positive and/or negative effect, and whether that effect acts on the extensive and/or intensive margin. I then assess whether each one is consistent with the negative extensive margin effects and positive intensive margin effects found in Section 4. I consider seven possible mechanisms below, which are summarized in Table 11.

i) Firstly, **formative learning** through testing could lead to positive testing effects; this can occur on both the extensive and intensive margins. However, the NAEP is a purely summative assessment, and teachers do not receive any feedback or information on student performance. This means that their teaching methods (with respect to the entire classroom)

should not react to NAEP testing, making a positive effect on the extensive margin unlikely.

A formative experience via the students' initiative is still possible if he or she queries a teacher about questions asked in the NAEP test afterwards. This would create a positive effect on the intensive margin—as a greater number of students are tested, more of them will individually ask their teachers about the NAEP questions. However, while this is consistent with the estimated positive intensive margin effects, this too seems implausible since it would require a decent number of NAEP-tested students to remember the test materials and actively seek teacher feedback.

ii) Secondly, treated schools may have their **resources stretched** as a result of having to host the NAEP. As resources are diverted away from other educational uses, this would have a negative effect on both the extensive margin (fixed resource costs of hosting NAEP) and intensive margin (variable resource costs).

For instance, participating schools provide a physical space for selected students to take the NAEP in. Since similar numbers of students are tested at selected schools regardless of size, this fixed resource cost has a negative effect on the extensive margin, consistent with the estimated results.

Outside coordinators employed by NAEP to administer the test ensure that teachers' schedules are unaffected. Nonetheless, participating students will lose some class time, since NAEP tests take approximately 90 minutes to administer. If it were a question of missed lesson time for the selected students, then this time resource cost would manifest as a negative effect on the intensive margin as a greater proportion of students get tested. However, this is inconsistent with the the evidence of positive effects on the intensive margin.

Nonetheless, lesson time may still be negatively affected on the extensive margin if NAEP testing has negative externalities on non-participating students. At larger sampled schools, there will be students not selected to take the NAEP, since the requisite number of students needed for the sample has been reached. Even though these non-participating students remain in regular classes, teachers may not give the planned lessons to them while a portion

of the class is taking the NAEP (perhaps thinking that doing otherwise would unfairly disadvantage the latter group). If lesson time is displaced in such a spillover manner, then all students at sampled schools—regardless of NAEP participation—lose learning time, creating a negative effect on the extensive margin. This too would be consistent with the estimated results.

iii) Thirdly, teachers or students may be responding to **perceived stakes** or consequences based on NAEP performance. In actuality, there are no stakes attached to the NAEP. Even though there are no *actual* stakes involved, the possibility of *perceived* stakes cannot be entirely discounted. Teachers who do not understand the purpose and sampling procedures of NAEP may mistakenly believe that their school has been singled out for particular scrutiny. Students and parents may have similar misconstrued views. Parents especially may question why their school in particular has been selected, pressuring school administrators to improve. In this case, the NAEP may spur teachers and students to expend more effort, bringing about positive effects on the extensive margin. However, such a result is inconsistent with the negative extensive margin estimates.

Moreover, I find this misconceptions argument unsatisfactory given the information provided to schools. Administrators at selected schools are informed well in advance of their selection and presumably must explain to teachers why some of their students are being pulled out of class for NAEP participation. Notes are sent to parents of selected students with information about NAEP and how it is distinct from the state-wide standardized tests. The Education Department distributes a host of informational materials for teachers, parents and students to reaffirm NAEP's importance. These materials also reassure them that NAEP results are kept confidential and not used for evaluation.

iv) Testing may lead to reduced **intrinsic motivation**, as students become less excited about learning after sitting for NAEP. By taking the NAEP test, students are primed to focus on performance goals instead of learning goals. When they are encouraged to do their best for the assessment, even with no stakes involved, students may perceive doing well on

tests as the ultimate aim of education. Since such behavioral changes affect NAEP-tested students individually, this would correspond to a negative effect on the intensive margin, which is inconsistent with the observed estimates.

v) The psychological effect most often cited is **test anxiety**: the fear of poor performance in the assessment. Even for no-stakes tests such as NAEP, test anxiety can still manifest itself. McDonald (2001) notes that "if at any stage of an evaluation we feel unprepared, unsure of our ability, or feel we have not performed to our best, we may experience feelings of unease, apprehension, distress or depression." The anxiety generated by taking tests such as NAEP can affect the process of learning afterwards (Tobias, 1992), thereby leading to changes in performance in state-wide standardized tests administered subsequently.

Initially, modest amounts of anxiety may improve performance, as students are spurred to put in more effort. This would correspond to positive effects on the intensive margin. However, as test anxiety increases, students can become impaired, leading to negative effects on the intensive margin. The positive effects on the intensive margin estimated earlier suggest that test anxiety has yet to reach the levels whereby negative effects start manifesting, and instead provides constructive stress by encouraging students.

vi) **Test fatigue** can take its toll on students in terms of physical and psychic costs of test-taking. As more students are tested, more of them become fatigued, resulting in negative effects on the intensive margin. However, this is inconsistent with the estimated results.

vii) Lastly, **learning by doing** is one possible channel through which NAEP's testing effect operates in a positive manner on the intensive margin, as estimated earlier. As students sitting for the NAEP respond to the questions, knowledge regarding the subject matter is reactivated in their minds. As students review these concepts through testing, practicing the reading or math skills involved, they strengthen their grasp of the material, irrespective of the lack of formative feedback. Students could also develop better test-taking strategies through the experience. Being placed in a test-taking environment lets students familiarize

themselves with such conditions, helping them to better cope with future tests conducted in similar settings. In many ways, taking the NAEP test is akin to taking a practice test in preparation for the subsequent state-wide standardized tests.

Having considered these seven possibilities, three mechanisms appear to affirm the evidence in Section 4. The negative treatment effects on the extensive margin probably arise from resources being stretched as schools host the NAEP sessions. The positive treatment effects on the intensive margin are likely the result of positive test anxiety spurring students to work harder, as well as benefits from learning by doing. Of course, the effect of NAEP testing on both the intensive and extensive margins may well be a combination of multiple mechanisms manifesting themselves simultaneously. However, without alternative measures of related outcome variables to distinguish between every channel, it is difficult to say definitively the degree of contribution from each one. On the whole though, it seems that the negative effects cancel out the positive ones, resulting in a net zero impact.

Skeptics may contend that the NAEP test is merely a one-time extra testing session that is not likely to matter, constituting a "low-dosage" treatment. However, I would argue that the NAEP test is not as trivial as it seems, especially in the eyes of students. The entire NAEP session is 90 minutes long, which is roughly half the time-length of typical sittings of state-wide standardized tests, and certainly much longer than the regular tests or quizzes given in class. The fact that its administration followed a regimented procedure—students are brought into a separate testing space with assigned seating, given official-looking test documents, and overseen by outside examiners invigilating the test—may have signaled to participants that the NAEP test is a serious and meaningful assessment. This systematic administration and official impression of the NAEP tests is especially relevant to any psychological effects generated by the test. Moreover, the fact that many of the treatment effects were estimated with statistically significance refutes this, showing that the "dose" was large enough to generate meaningful effects.

Critics may argue that the statistically insignificant coefficients estimated in certain spec-

ifications could just be the result of a lack of statistical power, as opposed to the treatment effects actually being near-zero. However, for numerous other estimates, I find statistically significant and near-zero-magnitude results. These precisely-estimated results lend credence to the argument that testing effects are actually small and near-zero, showing that this particular quasi-experimental identification strategy *can* yield statistically appreciable but small estimates. Thus, the conclusions regarding negligible testing effects drawn from the analysis should not be discounted as low-powered.

6 External Validity

The NAEP assessments resemble other real-world standardized tests in many ways. As mentioned before, the testing environment and administration procedures mimic those of state-wide standardized tests. The materials and booklets are designed in similar fashion, and the content areas covered are comparable to the curriculum typical of American schools. On the other hand, two characteristics of the NAEP potentially distinguish it from other assessments: that the NAEP has no stakes for students and that it is purely summative. However, upon closer examination, state-wide standardized tests—the sort most-often criticized by parents and the media—are not actually that different from NAEP along these two dimensions.

To investigate further, I conducted an email and telephone survey of state-level education departments in the fifty states and the District of Columbia between January and March of 2017. Respondents were education department officials within the bureaus or divisions responsible for assessments. 49 of the 51 education departments contacted responded to the survey after one or multiple attempts. The survey questionnaire focused on two aspects of state-wide standardized tests: 1) whether there are any stakes for the students (as opposed to the adults); and 2) whether the assessments play a summative or formative role.[19] I concentrate on stakes for students because I am interested in student outcomes, which can

[19]The exact wording of the questions can be found in Supplemental Appendix C.

be affected by stakes for students through influence on mechanisms including text anxiety and intrinsic motivation. Given the email / telephone format, the answers were open-response and no prompts were provided unless the respondent did not address the question posed, or they asked for clarification.

Regarding the first concern, it turns out that like the NAEP, many state-wide standardized tests also do not have stakes for students. 59% of the respondents (29 states) reported that their standardized tests involve no stakes imposed by the state on students at any grade level. In such states, the assessments are used for long-term planning purposes, as well as for accountability of the adults rather than the students. Of the 20 states that did report the existence of stakes for students, only one state (Minnesota) reported that their state-wide standardized tests has them at every grade level tested. The remaining 19 noted that stakes exist only at certain grade levels. Of these 19, 15 have stakes at the high school level (between grades 9 and 12) in the sense that state-wide standardized testing formed part of the graduation requirements.[20] More relevant to the external validity of the NAEP tests I examine (which are administered in grades 4 and 8), only 3 of the 19 states reporting stakes for students have them at the grade 8 level, while only 7 reported stakes in one of the elementary grade levels (between grades 3 and 5).

Even among this group of 20 states with stakes for students, many of the administrators interviewed emphasized the low stakes nature of the tests. 35% stated that standardized test results formed only a small part of the overall assessment criteria. A similar number of them (again 35%, but not necessarily the same ones) also mentioned that students had multiple avenues to make up for weak performance in tests through alternative activities such as retakes and teacher determinations. That these two caveats were repeatedly brought up by the officials unprompted—given the open-response nature of the survey format—shows the concern that they had about such stakes being mistakenly over-emphasized, and suggests

[20]Meeting the requirements either involves passing the typical end-of-year state-wide test administered in a particular grade, or passing a standardized end-of-course test administered in the term the student takes that course required for graduation.

that the stakes for students constitute what should be considered as low stakes.

Regarding the second concern, a vast majority of state-wide standardized tests are summative in nature, similar to the NAEP. 88% of respondents (43 states) described their state-wide standardized tests as being summative, as opposed to being formative. Of the 6 states that reported formative assessments, some mentioned systems through which teachers could access individual student results to tailor teaching in response. However, it was unclear to what extent such formative processes actually occurred. This impression was in fact formed from some of the (again, unprompted) responses of the "summative" group.

A full 49% of those reporting summative assessments noted that because these state-wide standardized tests were administered at the end of the year, teachers may find it difficult or unproductive to look up individual-student scores from the previous year to tailor their teaching. Instead (as some suggested), it would be more helpful to administer their own pre-tests at the beginning of the year for formative purposes. In fact, 21 states specifically brought up other pedagogical tools besides the summative state-wide standardized tests (e.g. classroom materials, smaller less-formal formative/interim assessments) that are better suited for formative needs.[21]

Overall, these survey answers indicate that the majority of states in the US administer state-wide standardized tests with no stakes for the students. Even for states with some form of stakes, they are often low in nature and/or limited to specific grade levels, usually skewed towards the high school grades. Furthermore, the responses suggest that the preeminent purpose of these state-wide standardized tests is summative assessment, conducted mainly to measure achievement trends and formulate aggregate policy. These similarities between the NAEP and state-wide standardized tests indicate that my results found using the NAEP as a quasi-experimental treatment are broadly generalizable.

[21] Hart et al. (2015) reach similar conclusions in their district-level survey.

7 Conclusion

Does testing adversely affect student learning outcomes? My findings suggest that overall, NAEP testing has no statistically meaningful effect on student performance in subsequent state-wide standardized tests. In fact, the 2013 results for grade 8 students suggest that administering one more marginal full-scale test may actually be slightly beneficial.

My results contrast with both correlational studies (which often find negative effects) and experimental studies (which usually report positive effects) in the previous testing literature. With regards to the correlational studies, their negative effects may be the result of omitted variable bias. Since NAEP testing is used as a quasi-experiment in this paper's identification strategy, the estimated treatment effects are unbiased, which may explain the contrasting findings.

With regards to previous experimental studies, there are some differences between them and this study. The NAEP is a purely summative assessment , whereas assessments administered in these other studies may have formative components to them. This perhaps highlights the importance of formative assessment, and its potential to compensate for and overcome any negative effects of testing. These other experimental studies are also consistent with my hypothesis that positive test anxiety and learning by doing are the two mechanisms driving the positive effects on the intensive margin. This is because their isolated experimental setups do not suffer from resources being stretched, as in the case of schools hosting NAEP.

Although my focus has been on the effects of testing on learning outcomes, there are numerous other outcomes of interest for possible future research. These include measures of test anxiety, psychological well-being, attitudes towards learning, and future labor market outcomes. Examining these can shed further light on the precise mechanisms behind the results found in this study. Furthermore, the NAEP treatment involves the addition of one test to students' schedule. If testing policies involved multiple assessments, it would be interesting to investigate what their cumulative effects are, and whether the marginal effect

of each additional test changes.

The central contributions of this paper are twofold. Firstly, my identification strategy provides internally valid estimates of the causal effect of testing. No other study utilizes the randomized nature of survey sampling design in this way. Secondly, these estimated effects provide greater external validity compared to previous experimental studies given the large-scale setting of the *natural* experiment involved, as well as the similarity of the NAEP treatment to other real-world standardized tests.

Understanding the effects of testing on educational outcomes is especially useful for policy-makers formulating standardized testing guidelines. Some may point to the necessity of balancing the trade-off between the effects of testing and its usefulness as a summative assessment tool for measuring educational progress. The estimates found in this paper imply that the negative effects of testing so often decried in the media are in fact negligible, at least when it comes to effects on subsequent learning outcomes. On the other hand, the collection of standardized test scores through summative assessment has been immensely beneficial to administrators for making effective policy decisions, as well as to parents for staying informed about the performance of their local schools. Given these findings, perhaps the trade-off is not as worrying as espoused.

References

Ames, Carole (1992) "Classrooms: Goals, Structures, and Student Motivation.," *Journal of Educational Psychology*, Vol. 84, No. 3, pp. 261–271.

Angrist, Joshua D. and Victor Lavy (1999) "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *The Quarterly Journal of Economics*, Vol. 114, No. 2, pp. 533–575.

Black, Paul and Dylan Wiliam (1998a) "Assessment and Classroom Learning," *Assessment in Education: Principles, Policy & Practice*, Vol. 5, No. 1, pp. 7–74.

——— (1998b) *Inside the Black Box: Raising Standards Through Classroom Assessment*: Granada Learning.

Carpenter, Shana K. and Jonathan W. Kelly (2012) "Tests enhance retention and transfer of spatial learning," *Psychonomic Bulletin & Review*, Vol. 19, No. 3, pp. 443–448.

Carpenter, Shana K., Harold Pashler, and Nicholas J. Cepeda (2009) "Using tests to enhance 8th grade students' retention of U.S. history facts," *Applied Cognitive Psychology*, Vol. 23, No. 6, pp. 760–771.

Crooks, Terence J. (1988) "The Impact of Classroom Evaluation Practices on Students," *Review of Educational Research*, Vol. 58, No. 4, pp. 438–481.

Firestone, William A., David Mayrowetz, and Janet Fairman (1998) "Performance-Based Assessment and Instructional Change: The Effects of Testing in Maine and Maryland," *Educational Evaluation and Policy Analysis*, Vol. 20, No. 2, pp. 95–113.

Gijbels, David, Gerard van de Watering, and Filip Dochy (2005) "Integrating Assessment Tasks in a Problem-based Learning Environment," *Assessment & Evaluation in Higher Education*, Vol. 30, No. 1, pp. 73–86.

Harlen, Wynne and Ruth Deakin Crick (2003) "Testing and Motivation for Learning," *Assessment in Education: Principles, Policy & Practice*, Vol. 10, No. 2, pp. 169–207.

Harris, Elizabeth A. (2015) "As Common Core Testing Is Ushered In, Parents and Students Opt Out," *The New York Times*, March.

Hart, Ray, Michael Casserly, Renata Uzzell, Moses Palacios, Amanda Corcoran, and Liz Spurgeon (2015) "Student Testing in America's Great City Schools: An Inventory and Preliminary Analysis," report, Council of the Great City Schools.

Heilig, Julian Vasquez and Linda Darling-Hammond (2008) "Accountability Texas-Style: The Progress and Learning of Urban Minority Students in a High-Stakes Testing Context," *Educational Evaluation and Policy Analysis*, Vol. 30, No. 2, pp. 75–110.

Hembree, Ray (1988) "Correlates, Causes, Effects, and Treatment of Test Anxiety," *Review of Educational Research*, Vol. 58, No. 1, pp. 47–77.

James, M. (2010) "An Overview of Educational Assessment," in Penelope McGaw, Eva Peterson, and Barry Baker eds. *International Encyclopedia of Education (Third Edition)*, Oxford: Elsevier, third edition edition, pp. 161–171.

Jones, M. Gail, Brett D. Jones, Belinda Hardin, Lisa Chapman, Tracie Yarbrough, and Marcia Davis (1999) "The Impact of High-Stakes Testing on Teachers and Students in North Carolina," *The Phi Delta Kappan*, Vol. 81, No. 3, pp. 199–203.

Kellaghan, Thomas, George F. Madaus, Anastasia E. Raczek, and American Educational Research Association (1996) *The Use of External Examinations to Improve Student Motivation*: American Educational Research Association.

Madaus, George F. and Marguerite Clarke (2001) "The Adverse Impact of High Stakes Testing on Minority Students: Evidence from 100 Years of Test Data," in Gary Orfield

and Mindy L. Kornhaber eds. *Raising Standards or Raising Barriers?: Inequality and High Stakes Testing in Public Education*: Century Foundation Press.

McDaniel, Mark A., Janis L. Anderson, Mary H. Derbish, and Nova Morrisette (2007) "Testing the testing effect in the classroom," *European Journal of Cognitive Psychology*, Vol. 19, No. 4-5, pp. 494–513.

McDonald, Angus S. (2001) "The Prevalence and Effects of Test Anxiety in School Children," *Educational Psychology*, Vol. 21, No. 1, pp. 89–101.

Mislevy, Robert J., Albert E. Beaton, Bruce Kaplan, and Kathleen M. Sheehan (1992) "Estimating Population Characteristics from Sparse Matrix Samples of Item Responses," *Journal of Educational Measurement*, Vol. 29, No. 2, pp. 133–161.

Niederle, Muriel and Lise Vesterlund (2010) "Explaining the Gender Gap in Math Test Scores: The Role of Competition," *The Journal of Economic Perspectives*, Vol. 24, No. 2, pp. 129–144.

Pennebaker, James W., Samuel D. Gosling, and Jason D. Ferrell (2013) "Daily Online Testing in Large Classes: Boosting College Performance while Reducing Achievement Gaps," *PLoS ONE*, Vol. 8, No. 11, p. e79774, 11.

Prodromou, Luke (1995) "The Backwash Effect: From Testing to Teaching," *ELT Journal*, Vol. 49, No. 1, pp. 13–25.

Roediger, Henry L., III and Jeffrey D. Karpicke (2006) "Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention," *Psychological Science*, Vol. 17, No. 3, pp. 249–255.

Roediger, Henry L., III and Elizabeth J. Marsh (2005) "The Positive and Negative Consequences of Multiple-Choice Testing," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 31, No. 5, pp. 1155–1159.

Santibanez, Lucrecia (2010) "Teacher Incentives," in Penelope Peterson, Eva Baker, and Barry McGaw eds. *International Encyclopedia of Education (Third Edition)*: Elsevier, pp. 481–488.

Tobias, Sigmund (1992) "The Impact of Test Anxiety on Cognition in School Learning," in K.A. Hagtvet and T.B. Johnsen eds. *Advances in Test Anxiety Research*, Vol. 7, Lisse, The Netherlands: Swets & Zeitlinger, pp. 18–31.

Yerkes, Robert M. and John D. Dodson (1908) "The relation of strength of stimulus to rapidity of habit-formation," *Journal of Comparative Neurology and Psychology*, Vol. 18, No. 5, pp. 459–482.

Zernike, Kate (2015) "Obama Administration Calls for Limits on Testing in Schools," *The New York Times*, October 24.

Supplemental Appendix

A Standardized Scores

This appendix describes the procedure used to create the school-level average standardized score variable. This procedure is necessary because results from statewide standardized tests are expressed not as average performance measures, but as a vector of percentages of students in separate ordinal performance levels (e.g. "advanced", "basic", "acceptable"). Further complicating this calculation is the fact that in the data: (a) each state has different numbers of performance levels ranging between 3 and 5; (b) each state separately defines what constitutes academic standards deserving classification into any performance level (e.g. different curricula and grading standards); and (c) data from 2010 and 2011 are partially censored. Regarding this last point, in 2010 and 2011, all states are observed to have only two overarching levels ("proficient" and "not proficient"), which are generated by the Education Department by mapping the various state performance levels into these two overarching levels.

The following notation will be used.

- Let y_{is} be the standardized score of student i at school s.
- Let N_s be the number of students at school s.
- Let y_s be the average standardized score at school s, given by

$$y_s = \frac{1}{N_s} \sum_i y_{is}$$

- Let Y be the random variable denoting the distribution of y_{is} for a particular state. Assume a normal distribution where $Y \sim N(\mu, \sigma^2)$.
- Let $\phi(.)$ and $\Phi(.)$ denote the standard normal pdf and cdf respectively.

- Let α_k be the standardized score cutoff for achieving performance level k, where higher k implies better academic performance. That is, if a student obtains a score in the interval $(\alpha_k, \alpha_{k+1}]$, then the student is classified into performance level k. Since the lowest performance level is $k = 1$, let $\alpha_1 = -\infty$. The total number of levels k and the location of these cutoffs vary across states.

- Let P_k be the proportion of all students in NAEP-treated schools classified into performance level k for a particular state. These values are observed in the data.

- Let p_{ks} be the proportion of all students in school s classified into performance level k. These values are observed in the data.

The standardized score y_{is} is a nationally-representative academic achievement score based on the results from NAEP. It is normalized to have a mean of 0 and a standard deviation of 1 over the national distribution of achievement at the student-level at NAEP-tested schools. The score is calculated separately for each year and each subject (math and reading). The steps below describe how the school-level average standardized score y_s is generated for one year and one subject.

A.1 Estimate μ and σ^2 for each state

To estimate the mean and variance of the standardized score y_{is} for each state, I use the scaled scores from NAEP. NAEP scaled scores are available separately for math and reading, and use their internal scoring system which ranges around 100 to 500 points.

First, the national-level (population) mean and variance of the NAEP scaled scores are calculated using the plausible values method based on Item Response Theory (see Mislevy et al. (1992)). Standard errors (and hence population variance) are estimated using the Jackknife method, which takes into account the survey structure of NAEP. Using similar methods, state-level (population) means and variances of the NAEP scaled scores are calculated for each state.

Next, using the calculated national-level mean and variance, the state-level means and variances of NAEP scaled scores are normalized such that they have a mean of 0 and a standard deviation of 1 over the national distribution of achievement at the student-level at NAEP-tested schools. Thus, the mean and variance for each state is given by

$$\mu = \frac{mean_{state} - mean_{national}}{variance_{national}}$$

$$\sigma^2 = \frac{variance_{state}}{variance_{national}}$$

These two parameters describe the distribution of y_{is} within each state, which is assumed to be normally distributed $N(\mu, \sigma^2)$. To check the normality assumption, I construct QQ plots of the NAEP scores by year, grade and subject. Figure A1 shows that these QQ plots are all close to the 45-degree line, indicating that the NAEP scores indeed fit the normal distribution well. The distribution of reading scores on the right-side panels appear left-skewed, but only slightly. Overall, the normality assumption seems reasonable. Since NAEP-tested schools are randomly sampled, even though the parameters μ and σ^2 are calculated over NAEP-treated schools only, the parameters apply to the score distribution across all schools, including those not selected for NAEP.

Nationally, the distribution of standardized score y_{is} has mean 0 and standard deviation 1, but at the state level, μ will vary depending on the state's average performance in NAEP, and σ^2 will vary depending on how spread out students are within a state in terms of academic achievement. Each state's μ and σ^2 is used to calculate the performance level cutoffs α_k in the next step. They are not used as measures of performance at each school in the third step; instead, results from statewide standardized test administered separately are used for that purpose. The remaining steps describe the procedure to be repeated for each state, year, and subject.

Figure A1: Check for Normality Assumption using QQ Plots

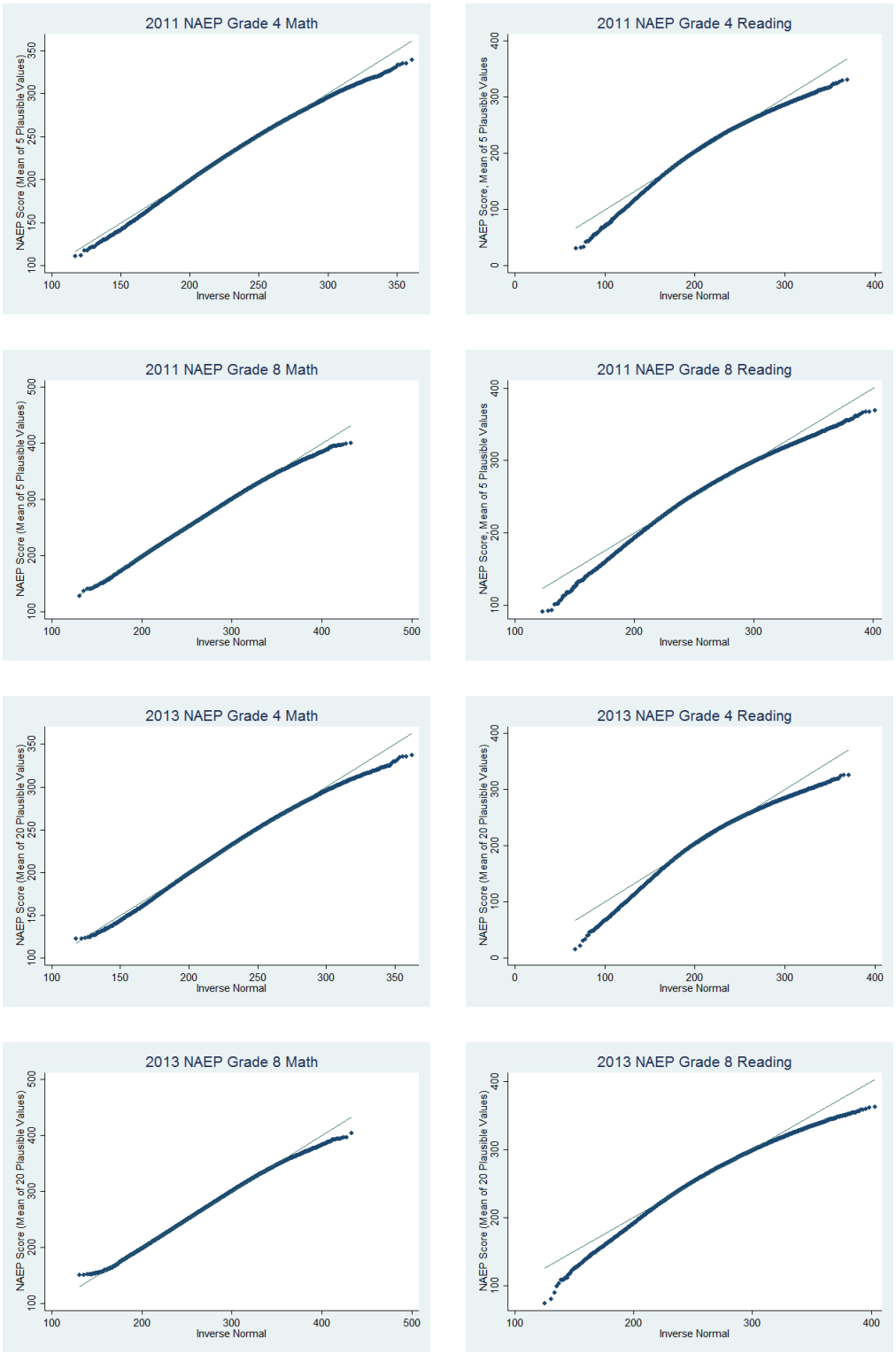
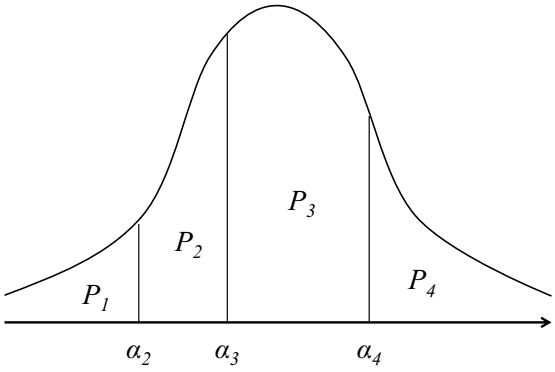


Figure A2: Cutoffs in the Standardized Score Distribution



A.2 Calculate performance level cutoffs α_k under normality

Suppose in a given state, there are K performance levels, indexed by $k = 1, \ldots, K$. From the data, calculate P_k, the proportion of all students in NAEP-tested schools classified into each performance level k.

Under the normality assumption, $Y \sim N(\mu, \sigma^2)$ in this state. Thus, each α_k can be calculated using the formula

$$\Phi\left(\frac{\alpha_k - \mu}{\sigma}\right) = 1 - \sum_{j=k}^{K} P_k$$

for $k = 2, \ldots, K$. This relationship is depicted in Figure A2. For $k = 1$, $\alpha_1 = -\infty$. For completeness in later formulas, define $\alpha_{K+1} = \infty$.

Even though these cutoffs are calculated based on the distribution at NAEP-treated schools, they are valid estimates of the cutoffs more generally applied across all schools, since NAEP-treated schools and control schools are not held to different statewide standards and are subject to the same cutoffs within a given state.

A.3 Calculate expected average school performance y_s under normality

Note that

$$E\left(y_s\right) = \frac{1}{N_s}\sum_i E\left(y_{is}\right)$$

Each individual student's score y_{is} is unobserved. Instead, only p_{ks}, the proportion of all students in school s classified into performance level k, is observed. Thus

$$E\left(y_s\right) = \sum_k \left[p_{ks} E\left(y_{is} \mid \alpha_k < y_{is} \leq \alpha_{k+1}\right)\right]$$

Under the normality assumption, the term $E\left(y_{is} \mid \alpha_k < y_{is} \leq \alpha_{k+1}\right)$ can be expressed using the inverse Mills ratio where

$$E\left(y_{is} \mid \alpha_k < y_{is} \leq \alpha_{k+1}\right) = \mu - \sigma\frac{\phi\left(\frac{\alpha_{k+1}-\mu}{\sigma}\right) - \phi\left(\frac{\alpha_k-\mu}{\sigma}\right)}{\Phi\left(\frac{\alpha_{k+1}-\mu}{\sigma}\right) - \Phi\left(\frac{\alpha_k-\mu}{\sigma}\right)}$$

Thus,

$$E\left(y_s\right) = \sum_k \left[p_{ks}\left(\mu - \sigma\frac{\phi\left(\frac{\alpha_{k+1}-\mu}{\sigma}\right) - \phi\left(\frac{\alpha_k-\mu}{\sigma}\right)}{\Phi\left(\frac{\alpha_{k+1}-\mu}{\sigma}\right) - \Phi\left(\frac{\alpha_k-\mu}{\sigma}\right)}\right)\right]$$

The expected school-level average standardized score is therefore a function of the proportions p_{ks} observed at that school, and state-specific parameters α_k, μ, and σ^2 as calculated from the previous steps.

For the special case where there are only two performance levels ($K = 2$), the formula is given by

$$\begin{aligned} E\left(y_s\right) &= \left(1 - p_{2s}\right)\left(\mu - \sigma\frac{\phi\left(\frac{\alpha_2-\mu}{\sigma}\right) - \phi\left(\frac{\alpha_1-\mu}{\sigma}\right)}{\Phi\left(\frac{\alpha_2-\mu}{\sigma}\right) - \Phi\left(\frac{\alpha_1-\mu}{\sigma}\right)}\right) + p_{2s}\left(\mu - \sigma\frac{\phi\left(\frac{\alpha_3-\mu}{\sigma}\right) - \phi\left(\frac{\alpha_2-\mu}{\sigma}\right)}{\Phi\left(\frac{\alpha_3-\mu}{\sigma}\right) - \Phi\left(\frac{\alpha_2-\mu}{\sigma}\right)}\right) \\ &= \left(1 - p_{2s}\right)\left(\mu - \sigma\frac{\phi\left(\frac{\alpha_2-\mu}{\sigma}\right)}{\Phi\left(\frac{\alpha_2-\mu}{\sigma}\right)}\right) + p_{2s}\left(\mu + \sigma\frac{\phi\left(\frac{\alpha_2-\mu}{\sigma}\right)}{1 - \Phi\left(\frac{\alpha_2-\mu}{\sigma}\right)}\right) \end{aligned}$$

with the only one defined cutoff α_2. This version of the formula will be applicable for the

2011 data, which only contain the percentages of students deemed "proficient" and "not proficient" at a school.

This formula is valid for both NAEP-treated schools and control schools, because normality is assumed for the score distribution across all schools. When $E\left(y_{is} \mid \alpha_k < y_{is} \leq \alpha_{k+1}\right)$ is substituted, it is the best guess of the expected standardized score y_{is} for performance level k based on information about the fixed distribution Y and the location of the cutoffs α_k applicable to both school types. This is independent of whether the school is NAEP-treated or part of the control group since NAEP-treated schools are randomly selected.

For example, suppose $E\left(y_{is} \mid \alpha_2 < y_{is} \leq \alpha_3\right) = 0.3$. This just means that if a student is observed to be in performance level 2, the best guess of his or her score is 0.3 standardized units. This best guess remains unchanged regardless of whether the student is from a NAEP-treated or control school. Information from the distribution at NAEP-treated schools is used only to determine the best guess of standardized scores in each performance level bin. This is separate from using the information from the school (p_{ks}) to calculate the average standardized score, a weighted average of performance level expected values and the proportion of students in each level at that school.

The same procedure of calculating school-level average standardized scores is used both for concurrent-year scores (statewide standardized tests taken after the administration of NAEP, e.g. 2013 grade 4), as well as for lagged scores (statewide standardized tests taken one year ago one grade below, e.g. 2012 grade 3).

A.4 Remark on Normalization Relative to Treatment Group

The NAEP scaled scores are used to normalize the distribution of achievement because the NAEP is comparable across states. The distribution of standardized scores is constructed to have mean 0 and standard deviation 1 over NAEP-tested schools only. The scores of the control schools are then fitted relative to this normalized distribution of the treated group.

This is opposite of what is usually done for normalizations in the treatment-effects lit-

erature. Usually, the control group's score distribution is normalized to have mean 0 and standard deviation 1, and the treatment group's scores are calculated relative to the control group's normalized distribution. Even though I do the normalization the other way around, this will not affect the interpretation of estimated treatment effects. The difference in test scores between treatment and control groups is measured in standard deviation units, so whether the normalization is done relative to the treatment group or the control group, the calculated magnitude of the difference is the same.

Consider the following example. First, suppose the normalization is done relative to the control group, such that the distribution of the control group is $N(0,1)$ and the distribution of the treatment group is $N(\beta,1)$, where β is a non-heterogeneous average treatment effect that has a constant effect across the entire distribution. The variances of the two distributions are the same because treatment and control group assignment is random, so on average, the two groups will look the same in that respect. Next, suppose the normalization is now done relative to the treatment group. Then the distribution of the control group will be $N(-\beta,1)$ while the distribution of the treatment group is $N(0,1)$. In both cases, the treatment effect will always be the difference in means of the two distributions, β.

B Robustness Checks

In this appendix, I perform robustness checks to confirm that the results presented in Tables 3 and 4 are robust to a variety of specification changes. This appendix considers four specifications with alternate sets of controls; all specifications include the treatment measures of interest, lagged standardized score and district fixed effects.

- Specification (A) is the original specification from before, which uses grade-level covariates where available, and school-level covariates otherwise.
- Specification (B) replaces these grade-level covariates with school-level ones, so all controls are calculated at the school-level.

Table B1: Summary of Controls Included in Various Regression Specifications

Specification:	(A) Original	(B) School -level	(C) School -level+	(D) Both Levels	(E) Minimal
Lagged Standardized Score	X	X	X	X	X
Race Proportions in Grade	X			X	
Race proportions at School		X	X	X	
% Boys in Grade	X			X	
% Boys at School		X	X	X	
Grade Size	X		X	X	X
School Size		X	X	X	
School % Free/Reduced Lunch	X	X	X	X	
Charter Indicator	X	X	X	X	
Local Dummies	X	X	X	X	
District Fixed Effects	X	X	X	X	X

- Specification (C) adds one additional control, number of students in the grade, to Specification (B). "School-level+" is used to denote this specification.

- Specification (D) includes the combined (union) set of grade- and school-level covariates from Specifications (A) and (B).

- Specification (E) includes the minimal controls needed for proper identification: lagged standardized score (so results can be interpreted as value-added) and the number of students in the grade (because sampling probability is proportional to grade size).

The exact set of controls in each specification is summarized in Table B1.

Table B2 presents robustness checks for the effect of NAEP testing on standardized scores, corresponding to Table 3 in the main paper. Table B3 presents robustness checks for the effects on the intensive and extensive margins, corresponding to Table 4 in the main paper. All alternate specifications considered show qualitatively similar coefficient estimates compared to the original specification. In particular, Specifications (C) through (E) have point estimates that are statistically indistinguishable from those of Specification (A).

Table B2: Robustness Checks for Effect of NAEP Testing on Standardized Scores

			(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
			Math				Reading			
			Grade 4		Grade 8		Grade 4		Grade 8	
	Specification	Variable	2011	2013	2011	2013	2011	2013	2011	2013
(A)	Original	NAEP-Tested	-0.0059***	-0.0001	-0.0020	0.0040	-0.0057***	-0.0058**	-0.0019	-0.0010
			(0.0021)	(0.0028)	(0.0039)	(0.0036)	(0.0016)	(0.0024)	(0.0029)	(0.0033)
(B)	School-level	NAEP-Tested	-0.0053**	0.0009	0.0011	0.0091***	-0.0056***	-0.0049**	-0.0003	0.0018
			(0.0021)	(0.0028)	(0.0038)	(0.0034)	(0.0016)	(0.0024)	(0.0029)	(0.0031)
(C)	School-level+	NAEP-Tested	-0.0064***	-0.0001	-0.0031	0.0043	-0.0060***	-0.0057**	-0.0028	-0.0009
			(0.0020)	(0.0028)	(0.0039)	(0.0035)	(0.0016)	(0.0024)	(0.0029)	(0.0032)
(D)	Both Levels	NAEP-Tested	-0.0062***	-0.0001	-0.0029	0.0038	-0.0061***	-0.0059**	-0.0027	-0.0008
			(0.0021)	(0.0028)	(0.0038)	(0.0035)	(0.0015)	(0.0024)	(0.0029)	(0.0031)
(E)	Minimal	NAEP-Tested	-0.0066***	0.0008	-0.0028	0.0036	-0.0064***	-0.0059**	-0.0036	-0.0027
			(0.0021)	(0.0028)	(0.0040)	(0.0036)	(0.0016)	(0.0025)	(0.0029)	(0.0033)

Significance Levels: *** = 1% ; ** = 5%; * = 10%.

Notes: Each panel labeled (A) through (E) presents a specification with a certain combination of controls summarized in Table B1. Within the panel, each row shows the coefficient estimate for the stated variable. Each column reports results for a separate regression specification with the dependent variable being school-average standardized scores for a specific subject, grade, and year. Standard errors clustered at the district level are in parentheses. All regressions include the "NAEP-Tested" treatment indicator, lagged standardized scores, controls, and district fixed effects.

Table B3: Robustness Checks for Intensive and Extensive Margin Effects

Specification	Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		Math				Reading			
		Grade 4		Grade 8		Grade 4		Grade 8	
		2011	2013	2011	2013	2011	2013	2011	2013
(A) Original	NAEP-Tested	-0.031***	-0.013	-0.014*	-0.014**	-0.016***	-0.015*	-0.009*	-0.020***
		(0.007)	(0.009)	(0.008)	(0.006)	(0.006)	(0.009)	(0.005)	(0.005)
	Treatment Intensity	0.031***	0.018	0.024*	0.045***	0.013*	0.013	0.014	0.045***
		(0.009)	(0.012)	(0.013)	(0.013)	(0.007)	(0.011)	(0.011)	(0.015)
(B) School-level	NAEP-Tested	-0.018**	-0.003	-0.001	0.0005	-0.011*	-0.007	-0.001	-0.011**
		(0.008)	(0.009)	(0.007)	(0.006)	(0.006)	(0.009)	(0.005)	(0.005)
	Treatment Intensity	0.015*	0.005	0.005	0.024*	0.006	0.003	0.001	0.031**
		(0.009)	(0.012)	(0.013)	(0.012)	(0.007)	(0.011)	(0.011)	(0.015)
(C) School-level+	NAEP-Tested	-0.031***	-0.016*	-0.015**	-0.014**	-0.015**	-0.018**	-0.008	-0.019***
		(0.007)	(0.009)	(0.007)	(0.006)	(0.006)	(0.009)	(0.005)	(0.005)
	Treatment Intensity	0.030***	0.021*	0.024*	0.044***	0.011	0.017	0.011	0.043***
		(0.008)	(0.012)	(0.013)	(0.013)	(0.007)	(0.011)	(0.011)	(0.015)
(D) Both Levels	NAEP-Tested	-0.029***	-0.014	-0.013*	-0.014**	-0.015**	-0.017**	-0.008	-0.018***
		(0.007)	(0.009)	(0.007)	(0.006)	(0.006)	(0.009)	(0.005)	(0.005)
	Treatment Intensity	0.028***	0.019	0.021	0.043***	0.011	0.015	0.011	0.043***
		(0.008)	(0.012)	(0.013)	(0.013)	(0.007)	(0.011)	(0.011)	(0.015)
(E) Minimal	NAEP-Tested	-0.030***	-0.001	-0.017**	-0.012**	-0.013**	0.0003	-0.011**	-0.019***
		(0.008)	(0.010)	(0.008)	(0.006)	(0.007)	(0.010)	(0.006)	(0.006)
	Treatment Intensity	0.029***	0.002	0.029**	0.039***	0.008	-0.008	0.016	0.040***
		(0.009)	(0.013)	(0.014)	(0.013)	(0.008)	(0.013)	(0.012)	(0.015)

Significance Levels: *** = 1% ; ** = 5%; * = 10%.

Notes: The same notes in Table B2 apply here. The only difference is that all regressions now also include treatment intensity as measured by proportion of students tested for NAEP.

The estimates using Specification (B) do change slightly, but there is a good reason for this. It is the only specification of the five presented that does not include grade size as a control. Including this covariate is essential for proper identification because for NAEP sampling, the probability of selection depends on the number of students in the grade. Though school size is highly correlated with grade size, including just school size as a control is clearly not enough. When grade size is re-included in Specification (C), the coefficient estimates revert to sizes similar to the original Specification (A).

Overall, these robustness checks confirm the validity of the estimates presented in the main paper, and strengthen the premise that the treatment effects are properly and causally identified using NAEP testing as the quasi-experimental framework.

C State-Level Education Departments Survey Questions

The following is the text of the survey questions administered to the state-level education departments. The same wording was used in the body of the emails sent as well as the oral script read out over the telephone.

1. Are there high stakes for the students taking the state-wide standardized assessments? That is, are there consequences for a student who does not perform well on them, such as being held back a grade or poor results showing up on a transcript that potential employers may see? I understand that there may often be high stakes for the teachers or schools, but I am interested in high stakes for the student in particular.

2. Are these state-wide standardized assessments formative in any way, or purely summative? That is, do individual students' results ever get reported to schools or teachers, and teachers can then adjust teaching in response to these results?